

Insights into DNA-protein interactions and functional clustering by studying motif distributions in the genome

विद्या वाचस्पति की
उपाधि की अपेक्षाओं की आंशिक पूर्ति में प्रस्तुत
शोध प्रबंध

A thesis submitted in partial fulfilment of the requirements of the
degree of Doctor of Philosophy

द्वारा / By

आत्रेयी चक्रबोर्ती / Atreyi Chakraborty

२०१९३६३९ / 20193639

डॉ. एम एस मधुसुधन / Dr. M.S Madhusudhan



भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान
पुणे

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

2024

*To
Aai, Baba, Ma and Suheli*

Declaration by Student

Name of Student: Atreyi Chakraborty

Reg. No.: 20193639

Thesis Supervisor(s): Prof. M.S Madhusudhan

Department: Biology

Date of joining program: 1st August 2019

Date of Pre-Synopsis Seminar: 19th April 2024

Title of Thesis: Insights into DNA-protein interactions and functional clustering by studying motif distributions in the genome

I declare that this written submission represents my idea in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I declare that I have acknowledged collaborative work and discussions wherever such work has been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

The work reported in this thesis is the original work done by me under the guidance of Dr./Prof. M.S Madhusudhan

Date: 27.09.24



Signature of the student

Certificate

Certified that the work incorporated in the thesis entitled “Insights into DNA-protein interactions and functional clustering by studying motif distributions in the genome” Submitted by Atreyi Chakraborty was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.



Date: 27.09.24

Prof. M.S Madhusudhan

Supervisor

IISER Pune

Acknowledgements

The completion of this thesis marks the culmination of a five-year journey that has been both challenging and rewarding in countless ways. As I reflect on these five years, I realize that this journey has been as much about personal growth as academic achievement. The challenges, the failures, and the moments of doubt were all made bearable by the incredible support I received along the way. This milestone would not have been possible without the unwavering support and encouragement of many people, to whom I owe a deep sense of gratitude.

I would like to express my deepest gratitude to my supervisor, Prof. Madhusudhan, whose mentorship went beyond academic instruction. Throughout these years, he has not only guided me in becoming a better researcher but has also taught me invaluable lessons about perseverance, integrity, and humility. Madhu always created a space where I could learn through my own mistakes, always encouraging me to explore ideas, no matter how unconventional, and to embrace the process of learning. He fostered a balance between providing direction and allowing me the freedom to grow independently. I am especially grateful for his patience, encouragement, and trust, which helped me navigate through the journey. Above all, he showed me, through their own actions, the importance of kindness, empathy, and respect in both science and life. I am incredibly fortunate to have had such a mentor, whose influence will continue to shape not just my career, but who I aspire to be as a person.

Special thanks to the COSPI lab members and colleagues for the many stimulating discussions, collaborative efforts, and moral support. With little background in coding when I started, I wouldn't have come this far without their guidance. They not only walked me through the basics but also challenged me to think critically and solve problems independently. A special mention goes to Sanjana Nair, who became my mentor and played a pivotal role in teaching me how to do experiments. I will always remember the long, tiring nights in the lab, troubleshooting and fine-tuning experiments together. Those late nights weren't just about getting results; they showed me the value of hard work, perseverance, and attention to detail. I am also incredibly grateful to Neelesh, Mukundan, Tejashree and Gulzar, who patiently taught me

programming and continually helped me improve my skills. They took the time to review my code, offer constructive feedback, and help me troubleshoot, even when the problems seemed insurmountable. Thank you for being instrumental in helping me grow both as a researcher and as a programmer. Tejashree was a great moral support, always there to offer advice and criticise in case I am falling behind. I am thankful to Tejashree, Shourya, Asita and Somanath for reviewing my thesis. I was lucky to collaborate with many talented individuals across different projects throughout these years. Each collaboration broadened my perspective, introduced me to new ideas, and pushed me to grow as a researcher. Working with Saswati, Sumant, Aparajita, Mrunal, Ajinkya, Saksham and Rohan, people from diverse backgrounds and skill sets enriched my experience and made the journey more fulfilling. A huge thanks to Aparna, for handling everything behind the scenes—from paperwork and travel grants to ensuring that everything was always well-organized. Her efficiency and attention to detail allowed me to focus on my research without worrying about the logistics. I truly appreciate their support throughout this journey.

I would like to express my deep sense of gratitude to Dr. Leelavati Narlikar, Dr. Kundan Sengupta and Dr. Durba Sengupta for providing me with the guidance, support and suggestions as a part of my research advisory committee. Their feedback during the annual RAC helped a lot with the progress of the work. I would also like to thank Dr. Richa Ricky for her insights and suggestions regarding the projects. I would like to thank Dr. Saikrishnan Kayarat and Dr. Gayathri Pananghat for allowing me access to their lab instruments and reagents for the experimental work. I am grateful for their willingness to help. I would also like to thank their lab members especially Apurba and Rushik for their help and suggestions during troubleshooting the experiments.

I would like to thank the IT team at IISER Pune Neeta Deo and Nisha Kurkure. Their expertise and prompt assistance in case of job errors or related to installation ensured the completion of tasks. I would like to thank the staff and members of the academic office and bio office for their help during this time. I would like to thank IISER Pune for the fellowship and DBT for providing me the travel grant.

I would like to extend my heartfelt gratitude to my high school teacher, Miss A. Banerjee who played a pivotal role in instilling my love for science. Her passion for the

subject and her ability to bring complex concepts to life sparked my curiosity and inspired me to explore the world of science further.

A heartfelt thanks to my friends Shaon and Nikita. Shaon was always there to help me and she was my first teacher when it came to solve coding problems. She was always there to help me with a smile on her face even after her long and tiring office hours. Nikita helped me with her suggestions and guidance regarding handling single cell omics datasets that made things easier for me. I'm extremely lucky to have friends like them.

I would like to thank Arnab, my biggest supporter, honest critic and ultimate partner in crime. His ability to motivate me when I'm feeling down and to challenge me when I need a push has shaped me into a better version of myself. His optimism has been my lifeline on the tough days. I genuinely don't know how I would have navigated this journey without him. I'm incredibly grateful to have him by my side.

Last but definitely not the least, I would like to express my deepest gratitude to my family—my parents and my sister—whose unwavering support and love have been the foundation of my journey. Ma and Baba, you have always believed in me, encouraging me to pursue my dreams, no matter how challenging they seemed. Their sacrifices and dedication to my education have inspired me to work hard and strive for excellence. Baba's endless optimism and Ma's principles were the guiding force all along. My sister, Suheli, was my confidante and cheerleader. Her understanding and encouragement during the toughest moments means the world to me. I would also like to honour my late grandmother, Aai who was the happiest when I joined this programme. I know she would have been overjoyed to see me reach this milestone and I carry her memory with me as I move forward in my career. I am incredibly fortunate to have such a loving and supportive family, and I dedicate this achievement to you all. I couldn't have done this without you.

I would like to extend my apologies if I've inadvertently missed anyone in this acknowledgment. So many people have contributed to this journey in ways big and small, and I am deeply grateful to all of you. Writing this acknowledgment was challenging, as my heart is full of so many memories and feelings. Each person who has supported me has played a unique role in shaping my experience, and it's difficult

to express my gratitude in just a few words. I am grateful for every lesson learned, every moment shared, and every act of kindness. Thank you to everyone who has been part of this journey; your impact on my life is immeasurable.

Table of Contents

Chapter 1: Overview	12
• Synopsis	13
• Thesis Organisation	16
 Chapter 2: Introduction to Techniques	 18
• Motifs and motif vector	19
• Extracting genome coordinates from experimental data	20
• Statistical measures	21
• Viterbi implementation of Hidden Markov Model	22
• Classification of Assessment measures	24
• CLICK	25
• JEDII (Juxtaposition Enabled DNA-binding Interface Identifier)	26
 Chapter 3: Reading Genomes	 28
• Optimal size of DNA motifs involved in specific interactions with proteins	33
• OE ratio highlights motif preference across chromosomes	34
• OE ratio clusters chromosomes based on motif preferences	37

• Motif distribution in the genome is non-random	39
Chapter 4: Promoter correlation to gene regulation	51
• Gene correlations from promoter motif distributions	55
• Significance of gene promoter correlations	59
• Features of gene promoters having high correlations	62
• Gene pairs with high correlation share common transcription factors	65
• Multiple gene regulation by a common transcription factor	69
• Promoter correlations highlight spatial proximity	70
Chapter 5: Motif - associated gene networks	80
• Finding motif abundances across genes	83
• Establishing gene associations for a single gene of interest	84
• Generating a network of multiple genes based on motif abundance	85
• Finding gene associations across chromosomes	86
Chapter 6: Role of motif clustering in chromosome positioning ...	88
• Distribution of LAD-iLAD regions across different chromosomes	95

• Identification of motifs enriched in LAD and iLAD regions	101
• Motifs enriched in LADs are AT rich	106
• Identification of evolutionarily conserved motifs enriched in LAD and iLAD regions	107
• Prediction of LAD-iLAD regions using Hidden Markov Model ...	109
• Identifying protein targets associated with LADs	111
Chapter 7: Conclusion and future prospects	115
• Summary of this study	116
• Compute database of gene network across multiple organisms	118
• Comparison of gene correlations across organisms	119
• Motif vector correlations to phylogeny	120
Data Availability	122
Publications	123
Bibliography	124
Copyright form	134

Chapter 1

Overview

- [Synopsis](#)
- [Thesis organisation](#)

1.1 Synopsis:

The genome sequence for any organism encodes the information for the regulation and maintenance of the biological machinery of the cell. The information encoded by the genes constitute a very small fraction (2%) of the entire genome. The majority of the genome is constituted of the non-coding regions. The regulation of the genome is brought about by the interaction of the regulatory binding partners. The information on the different binding targets from experimental data has a resolution of 10-100kbp. However, the optimal consensus binding consensus size is in the resolution of few base pairs. In this study we read the genome across the 24 different human chromosomes at the resolution of 5-6 bps, the resolution at which the specific recognition by the cognate protein binding partner occurs. We define these 5-6 bps stretches as DNA motifs. In this study we observed the motif distributions at the level of the whole genome as well as specific regions of the chromosomes. The findings highlighted distinct pattern of motif clustering at the different regions of interest (study was done for whole genome as well as small regions in the chromosome). We also established that the pattern of motif abundance and the motif clusters are distinct and non-random. On generating a random genome keeping the base composition same as original genome, the patterns are lost. This observation was consistent when studied at the level of the whole genome as well as smaller chromosome segments like the centromeric regions. Our analysis of the motif abundance in the centromeric regions could differentiate the acrocentric chromosomes from others. We employ the methodology of the motif abundances in two broad contexts: 1) Functional regulation of genes 2) Chromosome localisation.

For the first major goal of understanding the functional regulation of the genic regions in the chromosomes, we studied the motif abundances across the promoter regions (1,2 and 6kbps regions upstream of a gene transcription start site). Based on the motif abundances, we obtained the correlation score between all the annotated gene pairs in the human genome across all the different chromosomes. The analysis highlighted some significant findings. It was observed that the chromosomes involved in translocations share a higher percentage of the strongly correlated genes (Pearson's

correlation coefficient ≥ 0.9). We also established that the high correlations between the gene pairs are independent of the sequence similarity, i.e. a gene pair whose promoter shares an average of ~40% sequence identity also shares a high correlation. The observations for the correlations based on motif distribution in the promoter regions differ from those of the downstream regions of genes or random regions in the chromosomes. We validated the correlations obtained using three case studies. Three genes ADGRL1, TRIM8 and CDK9 present in chromosomes 19, 10 and 9 respectively that share a high correlation according to our analysis also showed regulation by common transcription factors between them. Our statistics of the motif abundance also highlighted the abundance of the consensus binding motifs significantly higher across the promoter regions of all the genes. Our findings highlight that correlations can be used to identify genes co-regulated by the common regulatory partner. For the second validation, we looked at an example of AP1 protein which consists of the Jun/Fos transcription factors that are reported to co-regulated multiple genes. We observed that among the 19 reported co-regulated genes, 10 out of the 19 genes share a high correlation between them from our analysis. The findings highlight that co regulated genes also share high correlation based on the motif abundances. For the third case study, we validated our gene correlations with the spatial proximity of genes using Hi-C and pcHi-C data. We obtained regions that are spatially proximal between chromosomes 18 and 19 which have distinct localization in the cell nucleus. We observed that the genes that share a high correlation between chromosomes 18 and 19 are more likely to be present in the spatially proximal regions. The high correlation between the genes is independent of the sequential proximity in the chromosomes.

We designed a web server, MAGNETIC, which has the database of all the gene pair correlations of all the 63,494 annotated genes in the latest T2T reference assembly. This count encompasses both protein-coding genes, pseudogenes, hypothetical genes and functional RNA (lncRNA, miRNA etc.) genes. The web server can be used to obtain all gene associations given a query gene and a network showing the associations chromosome-wise. The gene associations can be obtained for motif abundances at a resolution of 5 and 6 bases. The motif abundances can also be obtained for different promoter sizes of 1,2 and 6 kbps upstream of the gene transcription start site. The webserver can also be used to obtain a network of all the

gene associations given multiple gene lists. This network can be used to find the common connections between the genes at different cutoffs of gene pair correlations.

This organization of the genome inside the cell nucleus is a non-random event. For our second major goal we obtained the motif abundances in the context of the spatial localisation of the chromosomal segments and what maintains the positioning. The spatial positioning and localisation of the chromosomes are referred to as chromosome territories. The nuclear lamins, lining the nuclear envelope, have a large surface area to make potential contacts with the genome. These regions of high-frequency contacts between the chromosome and lamina are determined experimentally and are referred to as lamina-associated domains (LADs). The LADs constitute the chromosomal segments associated with the chromosome territories that have a peripheral organization. The experimentally determined LADs are in the size 0.5 to 1Mbps. We extended the analysis of finding the protein binding motifs of size 5 and 6 bases in these regions. Using experimentally available data, we identified the repetitive and conserved motifs across LADs in different chromosomes and those conserved in evolution. We identified the likely motif candidates responsible for the peripheral or interior tethering of the chromosomal segments in the cell nucleus. Using the motif distributions, we predicted the LAD and interLAD(iLAD) regions from genomic sequences by designing a Viterbi implementation of Hidden Markov Model. The work will give an insight into the functional significance of those motifs in DNA-protein interactions that are disrupted in laminopathies. We have also shortlisted the likely protein binding partners that can interact with and regulate the localization of the LADs based on the motifs obtained.

We have discussed the prospects of our method of using motif abundances to study different genomes in distinct prospects. We have also highlighted the preliminary data from our investigations of cross-genome correlations. The cross-genome correlations identify common motifs that have high abundances in the promoter regions across different genomes. This was done for 13 housekeeping genes in the human and mouse genomes and certain trends towards motif preferences were highlighted. Further elaborate investigations on the distribution of the motif abundances across genes in organisms are to be carried out. We also used this method of motif abundance to describe the evolution of 21 different strains of *S. cerevisiae*. These

results lead to further investigations on duplications of genome segments, how certain genome segments are translocated and the functional significance of the regions.

1.2 Thesis organisation:

1.2.1 Overview: Synopsis of the thesis and thesis organisation.

1.2.2 Introduction to techniques: A brief introduction to the various definitions and techniques used in the study. - motifs, handling experimental datasets, statistical measures, Viterbi implementation of Hidden Markov Model, classification of assessment measure, CLICK and JEDII.

1.2.3 Reading genomes: A detailed study of the human genome sequence at a resolution of motifs. Established why it is important to study the genome at the resolution of motifs. Quantified the motif abundance at the level of whole chromosomes as well as centromeric regions across all 24 human chromosomes. Established that the pattern of the motif distribution is non-random.

1.2.4 Promoter correlation to gene regulation: A study to establish a method to show associations between gene pairs based on motif distributions in the promoter regions. Obtained a correlation of all the annotated genes in the human genome based on the motif abundance in the promoter regions of the genes. The promoter regions were studied at different resolutions. Highlighted features of gene promoters which share high correlations. High correlations between genes were validated by common transcription factors shared between genes and the identification of the consensus sequence. The regulation of multiple genes by the same transcription factors was validated with our correlation scores. The high promoter correlations also showed spatial proximity.

1.2.5 Motif associated gene network: Description of a web server MAGNETIC (Motif Associated Gene NETwork In Chromosomes designed to access all the correlations between gene pairs across all the annotated human genes. The server

can be used to find gene associations for a single gene of interest or generate a network in case of multiple genes.

1.2.6 Role of motif clustering in chromosome positioning: A study highlighting the role of motif abundance in the context of chromosome positioning inside the cell nucleus. Detailed analysis of what motifs are abundant in the LAD vs the interLAD regions and how these motifs are evolutionarily conserved. A model was designed to predict LAD and iLAD regions from a given genome sequence using the Viterbi implementation of the hidden Markov model. Based on the motif information, likely protein targets that regulate the LAD and iLAD organisation were established.

1.2.7 Conclusions and future prospects: A summary of what has been done in the thesis and the prospects of the current study.

Chapter 2

Introduction to techniques

- [Motifs and motif vector](#)
- [Extracting genome coordinates from experimental data](#)
- [Statistical measures](#)
- [Viterbi implementation of Hidden Markov Model](#)
- [Classification of assessment measures](#)
- [CLICK](#)
- [JEDII \(Juxtaposition Enabled DNA-binding Interface Identifier\)](#)

2.1 Motifs and motif vector:

The genome sequence of any organism is made of long stretches of bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) arranged in a definite order which when packed inside the cell nucleus encodes the biological machinery of the cell. DNA sequence motifs are short, recurring patterns of DNA that are responsible for the function of cellular functions¹. These motifs are recognised by different proteins that enable their functioning. Protein binding partners (transcription factors/ regulators) bind to its cognate DNA sequences at a resolution of 5-6 bases². We have defined motifs as different combinations of 5-6 lettered words that can be recognized by a protein specifically (for example by hydrogen bonding). In our study, we use the terms k-mer, forward-reverse complement pairs and n-mers to classify different motif representations used in our analysis. All possible combinations of motifs (4^k where k is the size of the motif) are defined as the motif vector. We have categorized the motif representation into 3 further subsets ([figure 2.1](#)):

- K-mer:** The 5/6 base pair combinations representation of motifs.
- Forward-reverse complement pair:** We identify motifs as a pair considering a double-stranded DNA. We consider the forward reverse complement of a motif as a single representation.
- N-mer (or variable representation):** We introduce variable positions in the motifs by replacing A/T/G/C with N. N-mer representation will be a general representation of the different base combinations at the position replaced by N.

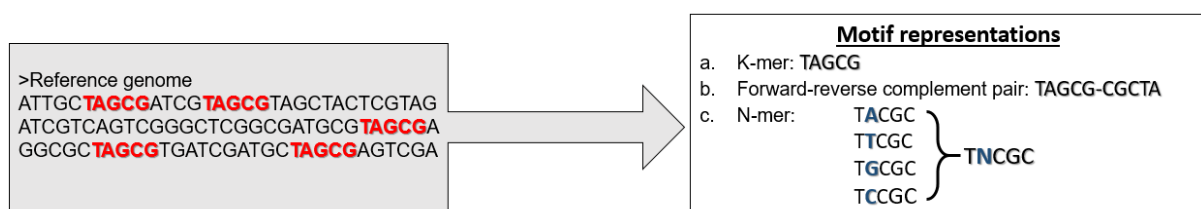


Figure 2.1: Schematic representation showing reading the genome sequence in terms of motifs (k-mer, forward and reverse complement pair and n-mer). The three representations of motifs used in the study are highlighted in the box on the right.

2.2 Extracting genome coordinates from experimental data:

We used the genome coordinates from different single-cell omics datasets to extract the sequence information for our calculation. The different experimental datasets used in our study are DamID data³ (interaction of genome segment with Lamin B), Hi-C⁴ and promoter capture(pc) Hi-C data⁵. For the coordinates of the spatially proximal genome segment, we used the .mcool files from Hi-C experiments. The .mcool files store the information of the Hi-C contacts at different genome bins in a compressed matrix format. The x-axis of the matrix stores the bin-wise coordinates (bins range from 10kbps to 250kbps depending on the resolution at which the experimental data was recorded). The y-axis records the contact frequency or the confidence score for the contacts made in the given range ([figure 2.2](#)). The higher the peak more likely the chances of an interaction observed in the given genomic range.

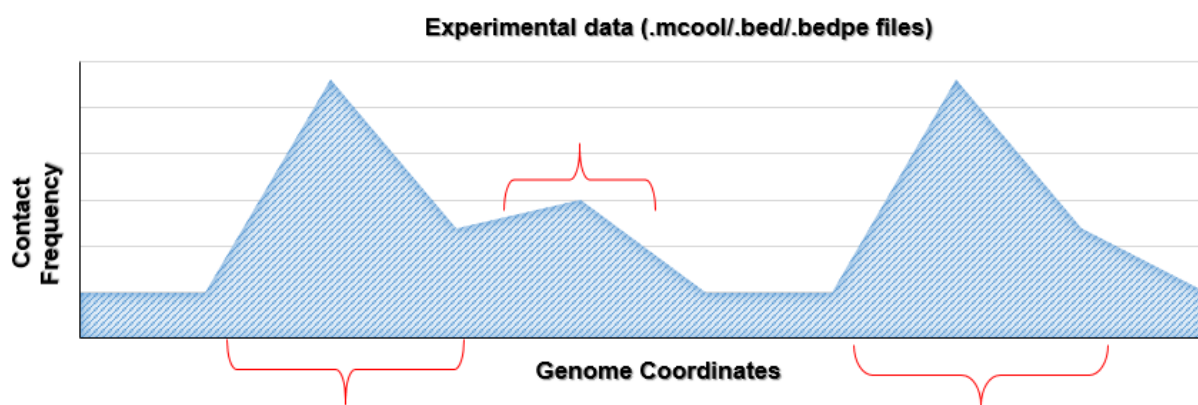


Figure 2.2: Extracting genome coordinates with interactions from DamID/ Hi-C/ pHi-C data. The y-axis represents the contact frequency and the x-axis represents the genome coordinates. The ranges highlighted in red correspond to the genome coordinates which were extracted for further analysis.

2.3 Statistical measures:

To compute the abundances of motifs and comparison of motif distribution across different regions we used the following different statistical measures:

2.3.1 Observed by expected ratio(OE): To analyse the motif distribution in the genome (or specific segment of interest in the genome) we used a statistical measure of the observed to expected (OE) ratio. The number of times a motif occurs in a sequence stretch is recorded as the observed count (O). The relative abundance is normalized by the expected count (E) which takes into account the AT/GC richness in the genome segment. E is calculated as $(N - k + 1) \prod_x f_x^{n_x}$ where n_x is the number of times a particular base occurs in a motif and whose probability is f_x and N is the size of the chromosome. The OE ratio captures the relative abundance of the motifs. OE ratio >1 signifies the occurrence of a motif higher than expected by chance and similarly, OE < 1 signifies the occurrence of a motif less than expected by chance. The OE ratio is captured across different ranges and resolutions of data. The observed count varies according to the chromosome segment studied, whereas the expected count is the same for an individual chromosome of interest.

2.3.2 LAD to iLAD ratio (LIR): The nuclear envelope is lined by the nuclear lamina. Regions in the chromosome which make frequent contact with the nuclear lamina are called the Lamina associated domains (LADs) and the interspersing chromosomal segments are called the interLADs (iLADs). We calculated the occurrences of all possible motif combinations from the motif vector in these regions. The position index of all the individual motifs in the motif vector was obtained using a pattern searching algorithm (naive string search) and stored in a .json file (dictionary with key as motif and list of positions as values) for each chromosome. To obtain the abundances of the motif in the LAD to iLAD regions we used a metric called the LAD to interLAD ratio (LIR). LIR is the normalized count of motifs in LAD to the interLAD region. The normalization was done according to the LAD coverage at the given contact frequency cutoff.

$$LAD \text{ to } iLAD \text{ ratio}(LIR) = \frac{\text{Normalized motif count in LAD}}{\text{Normalized motif count in iLAD}}$$

$$\text{Normalized count} = \frac{\text{Count in LAD/iLAD}}{\text{LAD/iLAD length in Mbps}}$$

An LIR greater than 1 signified a higher abundance of a motif of interest in the LAD region compared to the iLAD regions. Similarly, LIR less than 1 indicates a higher occurrence of motifs in the inter-LAD when compared to LAD regions.

2.3.3 Correlating motif vectors: We used correlation to compare the similarity between two motif vectors for different genomic regions. The correlation between two motif vectors, X and Y, is computed as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y} \text{ where } cov(X,Y) = \sum (X - \mu_X) \times (Y - \mu_Y) \quad (1)$$

Where $cov(X,Y)$ is the covariance, $(\mu_X$ and $\mu_Y)$ and $(\sigma_X$ and $\sigma_Y)$ are the means and standard deviations of OEs of motif vectors X and Y respectively. The correlation score ranges from -1 to +1 signifying the degree of correlation.

2.4 Viterbi implementation of Hidden Markov

Model:

A Hidden Markov Model (HMM) is a statistical framework used to represent systems with unobservable (hidden) states, characterized by a set of states, a set of observable events (emissions), transition probabilities (TP) between states, emission probabilities (EP) of the observable events given the states, and initial state probabilities (starting probabilities SP)⁶. The Viterbi algorithm is an efficient dynamic programming method used to determine the most probable sequence of hidden states in an HMM, given a sequence of observed events⁷. The algorithm initializes by setting up the initial probabilities for the first observation, then recursively calculates the highest probability of each state at each time step by considering the transition probabilities from previous states and the emission probabilities of the current observation. This process involves updating a matrix to keep track of the maximum probabilities, ultimately identifying the highest probability at the final time step. The algorithm then backtracks through the matrix to reconstruct the most likely path of hidden states that led to the observed

sequence, ensuring accurate decoding of the observations in terms of the underlying hidden states transitions ([figure 2.3](#)).

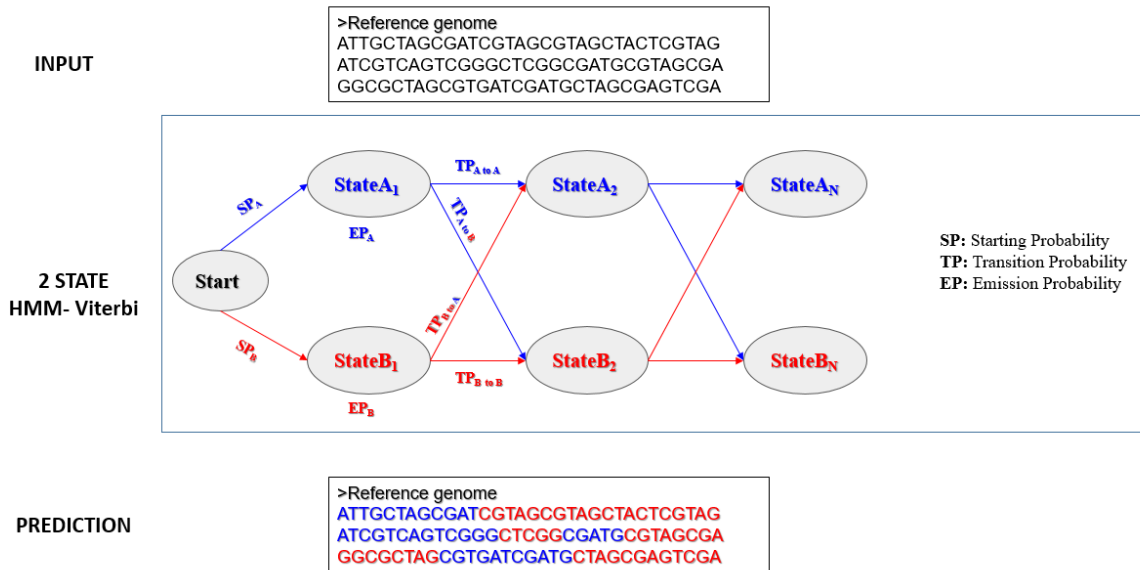


Figure 2.3: Schematic representation of the working HMM with Viterbi implementation. The input is a reference genome sequence shown in the first box. The second box highlights the working of the model using the different probabilities (SP, TP and EP) for each state shown in the flowchart. The prediction of the states based on the probabilities is highlighted in the last box.

The main recursive formula for the Viterbi algorithm is given by:

$$\delta(j) = \max_i [\delta(i) \times a_{ij}] \times b_j \times (O)$$

where:

- $\delta(j)$ is the highest probability of any path that ends in state j.
- $\delta(i)$ is the highest probability of any path that ends in state i.
- a_{ij} is the transition probability from state i to state j.
- $b_j \times (O)$ is the emission probability of the observation from state j

We designed a two-state Hidden Markov model to predict and distinguish a LAD from an interLAD region given a genomic sequence of interest. To find the most likely sequence of hidden states, we employed the Viterbi algorithm. It recursively computes the probability of each possible next hidden state. The calculation is based on the probability of the current hidden state and the transition probabilities between hidden

states. The algorithm generates a matrix of probabilities for each set of observations in a hidden state. We then select the hidden state with the highest probability and generate our prediction results.

2.5 Classification of assessment measures:

A classification scheme classifies data points into different categories. A binary classifier classifies data points into two categories. The performance of the binary classifier can be assessed using different metrics. Following are some of the widely used assessment metrics:

2.5.1 Confusion Matrix: Confusion matrix categorizes the binary classification predictions into four categories:

- True Positive (TP): The outcome where the model correctly predicts the positive class.
- False Positive (FP): The outcome where the model incorrectly predicts the positive class.
- True Negative (TN): The outcome where the model correctly predicts the negative class.
- False Negative (FN): The outcome where the model incorrectly predicts the negative class.

2.5.2 Sensitivity: It is a metric to measure the proportion of actual positive cases that the model correctly identifies. Sensitivity is calculated as the ratio of true positive predictions to the total number of actual positive cases, indicating the model's ability to accurately detect positive instances. It is also referred to as the true positive rate (TPR) and is calculated as:

$$Sensitivity (TPR) = \frac{TP}{TP + FN}$$

2.5.3 Specificity: It is a metric to measure the proportion of actual negative cases that the model correctly identifies. Specificity is calculated as the ratio of true negative predictions to the total number of actual negative cases, reflecting the model's ability to accurately detect negative instances. It is also referred to as the true negative rate (TNR) and is calculated as:

$$Specificity (TNR) = \frac{TN}{TN + FP}$$

2.5.4 Mathews correlation Coefficient: The Matthews Correlation Coefficient (MCC) is a metric used to evaluate the performance of a binary classification model. It takes into account all four categories (true positives, false positives, true negatives, and false negatives). It is a balanced measure when that is independent of the size of the classifiers. The values range from -1 to +1. The higher the positive value stronger the agreement of the model with the dataset. It is calculated as:

$$MCC = \frac{(TP + TN) - (FP + FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

2.5.5 Accuracy: It is a metric to measure the proportion of correct predictions (including both true positives and true negatives) relative to the total number of predictions. It is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2.5.6 Precision: It is a metric that quantifies the proportion of correct positive predictions. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

2.5.7 F1 score: The F1 Score is the harmonic mean of precision and recall (sensitivity), giving equal weight to both metrics. It is calculated as:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

2.6 CLICK:

Studying the similarity and overlap of 3D protein structures can provide valuable insights into a protein's function. CLICK^{8,9} is a computational tool that determines the protein structural overlap without being dependent on the topology of the proteins. It

is used to match the three-dimensional structures of proteins. CLICK, can match 3-D structures of proteins by comparing two sets of points regardless of their chain connectivity. CLICK creates small groups of points (3-7 in number) from selected atoms (as defined by the user, such as C α atoms or a combination of C α , C β atoms, etc.) of spatially close amino acid residues. These groups are then superimposed using a 3D least squares fit method. Additional features like solvent accessibility, secondary structure, and residue depth can also be used to guide the matching process.

2.7 JEDII (Juxtaposition Enabled DNA-binding Interface Identifier):

JEDII is a prediction algorithm for DNA-protein binding and relies on the similarity of hydrogen bond patterns at DNA-protein interfaces². It uses hydrogen bond donor and acceptor atoms in protein residues to guide structural alignments using the CLICK tool. The algorithm is based on the conservation of hydrogen bond donors and acceptors, which are crucial for DNA-protein specificity. By leveraging this conservation, the algorithm predicts the DNA sequence that binds to a given protein structure, based on the complementarity of the interface regions between the query protein and the template ([figure 2.4](#)).

In constructing the model, the DNA coordinates from the template are transferred to the aligned query. To ensure model accuracy, any models with more than one clash—defined as a distance less than 1.5 Å between protein C, C α atoms, and DNA atoms—are discarded. Main chain O and N atoms are not checked for clashes since they function as hydrogen bond acceptors and donors, respectively. The clash criterion is set at 1.5 Å, while hydrogen bond donor-acceptor distances below 4 Å are permitted. A DNA-protein contact is defined as a distance less than 3.5 Å between donor/acceptor atoms, and models with fewer than three such contacts are removed.

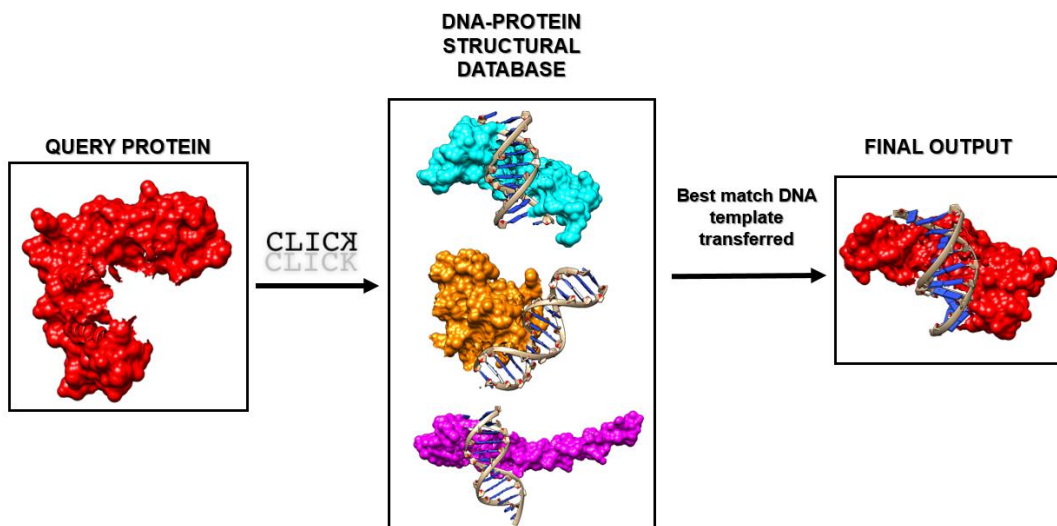


Figure 2.4: Schematic diagram of the JEDII algorithm.

The best model is selected based on the query overlap percentage, calculated as the number of donor and acceptor matches between the query and template proteins relative to the total number of donors and acceptors in the query protein. DNA-binding residues are identified as amino acids within 4 Å of any DNA atom, and the DNA sequence from the template is predicted as the binding sequence for the query protein. This methodology, named Juxtaposition DNA-binding Interface Identifier (JEDII), has been validated using the PDNA-285 dataset and other benchmark datasets².

Chapter 3

Reading Genomes

- [Optimal size of DNA motifs involved in specific interactions with proteins](#)
- [OE ratio highlights motif preference across chromosomes](#)
- [OE ratio clusters chromosomes based on motif preferences](#)
- [Motif distribution in the genome is non-random](#)

3.1 Introduction:

The human nuclear genome comprises a vast sequence of 3.2 billion base pairs of DNA, constituting 24 distinct chromosomes. The sequence encoded in the chromosomes regulates the functioning of the biological machinery of the cell. The latest T2T-CHM13 reference ¹⁰, has comprehensively annotated the whole genome without any gaps and has identified a total of 63,494 genes within the human genome. This total count of the genes is inclusive of all the protein-coding genes, pseudogenes, hypothetical genes and functional RNA (lncRNA, miRNA etc.) genes¹¹. It is important to note that the regions constituted by the genes only cover a small part of the genome with the non-coding regions covering the majority of the genome^{12–14}. These non-coding parts of the genome comprises of repetitive DNA regions, such as transposable elements¹⁵, consisting valuable information on disease-causing mutations¹⁶, genetic variations ^{17,18} and evolutionary conservation ^{19,11}. One major limitation of the previous human reference genome assemblies (GRCh38 and 37) was the unresolved sequence of 11.5 million bases which were uncharacterized and represented as gaps. With the advancement of sequencing methodologies, we have the complete human genome sequence. The T2TCHM13 genome assembly replaces the missing elements, updates the incomplete sequences, and rectifies structural elements, thus unravelling the complexities of the human genome. This additional information, amounting to approximately 200 million base pairs, will greatly enhance the precision and refinement of genome analysis and functional studies ^{11,18,19}. The current genome assembly has refined the annotations of specific genomic regions highlighting the understanding of segmental duplications and epigenetic markers^{20,21}.

The regulation of the transcriptome inside the nucleus is complex and dependant on the spatial organization²². Different experimental techniques are used to investigate the genome sequence to understand the spatial coordination and how the transcriptome is maintained^{23,24,25}. Different experimental techniques such as 3C techniques ⁴, Hi-C ^{18,26}, Dam ID ^{3,27}, Chip-Seq^{28,29}, ATAC Seq^{30,31}, super-resolution imaging³² and integrative genome modelling^{33,34} investigate these characteristics of spatial organization, protein-genome interaction and chromatin accessibility. The resolution of the information from the techniques mentioned above is usually in the range of 10-100 kilo-base pairs(kbps). Analysing genomes for all of the characteristics

mentioned above usually involves aligning sequence stretches to one another. However, relying solely on sequence alignment to investigate genome stretches of size 10-100 kbps, especially with the intent of gaining insights into genome organization/function has limitations. The small sample size of bases in such comparisons leads to a high signal-to-noise ratio, rendering DNA sequence alignment inconclusive. Even randomly generated sequences exhibit a similarity of approximately 25% or more. Therefore, identifying and comparing sequence patterns necessitates a focus on shorter sequence sizes. In this study we used an alternative alignment free method to ‘read’ the genome.

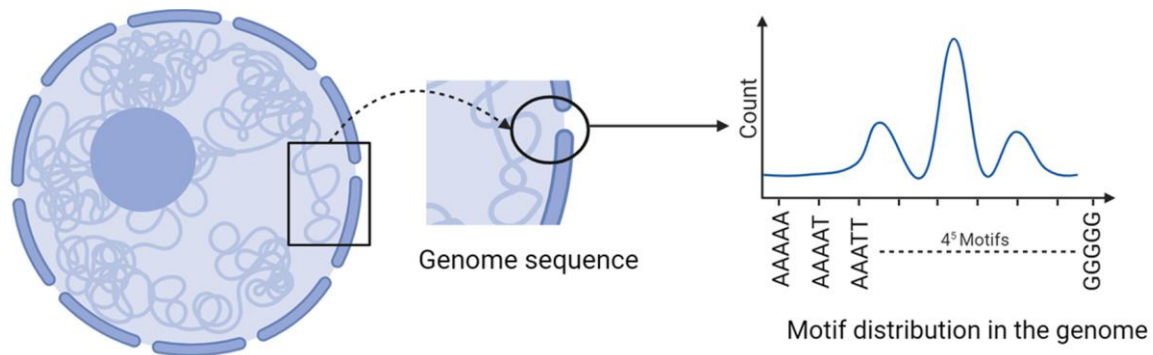


Figure 3.1: Schematic representation showing reading the genome sequence in terms of motifs.

Genomic studies have demonstrated that the periodicity of large genomes, spanning megabases (Mbps) in size, can be discerned by analysing trends in much smaller sequences, such as dinucleotide repeats or single sequence repeats found in microsatellites and minisatellites. Additionally, CpG islands, which play a role in regulating gene expression patterns, and trinucleotide single sequence repeats have been identified as important sequence motifs^{35,36}. Numerous other examples of repeat regions, particularly in the non-coding portion of the genome, have been reported. As the majority of the genome falls into this category, it is imperative to examine the resolution of DNA motifs to understand their conservation and determine how the frequency of occurrence correlates with their genomic locations.

K-mer or word frequency estimation are alternatives of alignment based studies where sequences are processed in moving windows of a specified word length^{37,38}. This method does not take into consideration the positional information of bases in

sequence. Nonetheless, it is valuable for comparison of long sequence stretches (in order of ~10-100 kbps) using calculation of word density to determine the similarity or dissimilarity^{39,40}. These methods assess sequence patterns without relying on alignments. In this study, we modified the k-mer method for a functional study of the genome.

DNA-bound protein complexes in PDB contact DNA specifically spanning a resolution of 5-6 nucleotides². There are cases of larger DNA sequence recognition, but these account for the cases where the binding protein partner oligomerizes and hence contiguous patches span a larger sequence stretch. There are other cases where contacts are made at different surface patches of the interacting protein. Only 5-6 bases (DNA motifs) are involved in base-specific recognition by hydrogen bonding. Hence, in this study, we characterized the genome at a resolution at which it will be specifically identified by the regulatory protein. Sometimes when protein interact with DNA, the DNA could bend or alter its conformation from the regular B-form^{41,42}. The flanking DNA sequence is also a key determinant of the affinity of the transcription factor^{43,44}. That notwithstanding, in this study we are assuming that a particular DNA motif when binding to its cognate protein would always undergo the same recognition shape change. We conducted an *in-silico* analysis to investigate the representation of the human genome in terms of DNA motifs. We analysed motif distribution across different sizes ranging from 2 to 6 bases. We compared the distinction between the human genome and a random genome as a control. The analysis was done for the whole genome as well as smaller genomic stretches such as the centromeric regions.

3.2 Methodology:

Data of genomic sequence(s) and the gene annotation, for all calculations was obtained from NCBI RefSeq Human Genome assembly T2T-CHM13v2.0.^{45,46}

3.2.1 Motif generation and distribution metric scoring:

To investigate the distribution of DNA motifs throughout the genome, we employed a sampling approach at a resolution of a few bases ([figure 3.1](#)). To assess the enrichment of specific motifs within certain regions, we employed the OE ratio (observed-to-expected ratio). The OE ratio is calculated as:

Observed count (O): Number of times a motif occurs in a genomic stretch

$$\text{Expected count (E)} = (N - k + 1) \prod_x f_x^{n_x}$$

$$\text{OE Ratio} = \frac{\text{Observed count}}{\text{Expected count}}$$

Where:

n_x : number of times a base occurs in a motif

f_x : base probability in the genome

k : motif length (in base pairs)

N : chromosome length (in base pairs)

The OE ratio is a statistical method used to measure the extent to which a particular feature is overrepresented or underrepresented within a given dataset. It is calculated by dividing the observed frequency of the feature by the expected frequency based on the probability of occurrence. It is a normalized metric that accounts for the AT/GC content of a chromosome. We calculate the OE ratio for all 4^k motif combinations possible given that there are 4 bases (A/T/G/C). The OE values are averaged for a given motif and its reverse complement. The motif vector is the OE count recorded for all combinations of motifs and their reverse complement pair. By calculating the observed frequency of motifs within a selected region and comparing it to the expected frequency, we aimed to gain insights into the distribution patterns of motifs across different regions of the chromosomes. This analysis provides valuable information regarding the enrichment of certain motifs in specific genomic regions.

Based on the genomic sequence range in this study, we obtained the observed count of the motifs and computed the OE ratios as follows:

- **OE_{whole}**: The first measure of the O/E ratio was done for the whole genome, where the observed count was the count of all the k-mer motif vectors in the whole genome.
- **OE_{100kbp}**: To examine the distribution on a local scale, we partitioned the chromosome sequence into bins. This bin-wise distribution provided us with increased degrees of freedom for analysis compared to the unbinned distribution. To segregate the chromosome sequences we divided the chromosome into bins of 100kbps. The observed count is the count of occurrence of a motif in a 100kbps bin.

- **OE_{centromere}:** To study the distribution patterns of motifs, in the centromeric region, we applied the same O/E ratio to compare the motif distribution in the centromeric regions of all chromosomes. The boundaries of these centromeric regions are well-defined from the NCBI genome data viewer ⁴⁶ (table 1), allowing for precise analysis. The observed and expected count is the count of occurrences and the probabilistic count of occurrences in the centromeric region.
- **OE_{random}:** We generated a randomized genome (explained in detail in section 3.2.2) recorded the observed count of all the k-mer in the motif vector at different genomic sizes and calculated the OE ratio.

In the above-mentioned calculations, the observed count is obtained from the different sequence stretches. The expected count is the same across all computations (for a particular chromosome).

3.2.2 Generating random genome:

As a control set, we generated a random genome for all 24 **distinct** chromosomes having the same length and base composition as the original sequence. We used the random module of python for the randomization. We used the nucleotide probability (or base frequency) f_x (where $x = A/T/G/C$) and the chromosome length from the original sequence. The randomization was based on selecting the nucleotide in a weighted random order. We also generated random set keeping the count of the nucleotides same as the original chromosome sequence. The randomization was done in triplicates to ensure consistency in comparisons.

3.3 Results:

3.3.1 Optimal size of DNA motifs involved in specific interactions with proteins:

We analysed the optimal DNA binding patch for specific recognition by proteins. We calculated the distance between two consecutive hydrogen-bonded nucleotides in an NR set of 285 DNA-bound protein complexes (the duplicate structures removed to avoid overrepresentation). From the frequency distribution of the distances, we

observed that the optimal size for the interaction lies below 6 bps and the number decreases with an increase in the distance ([figure 3.2a](#)). Cases where the distance between two consecutive hydrogen-bonded bases is more than 6bps, it is indicative of different binding patches. We observed that the hydrogen bonds that are separated by a distance more than 6 bps interact with amino acids of two distinct patches on the protein surfaces shown in purple in figure 3.2b. Thus, this observation highlights that when considering specific recognition by hydrogen bonding, the interaction does not encompass more than 5-6 bps stretch. Hence, we further characterize the genome at a resolution of these DNA motifs which will be identified by the regulatory proteins.

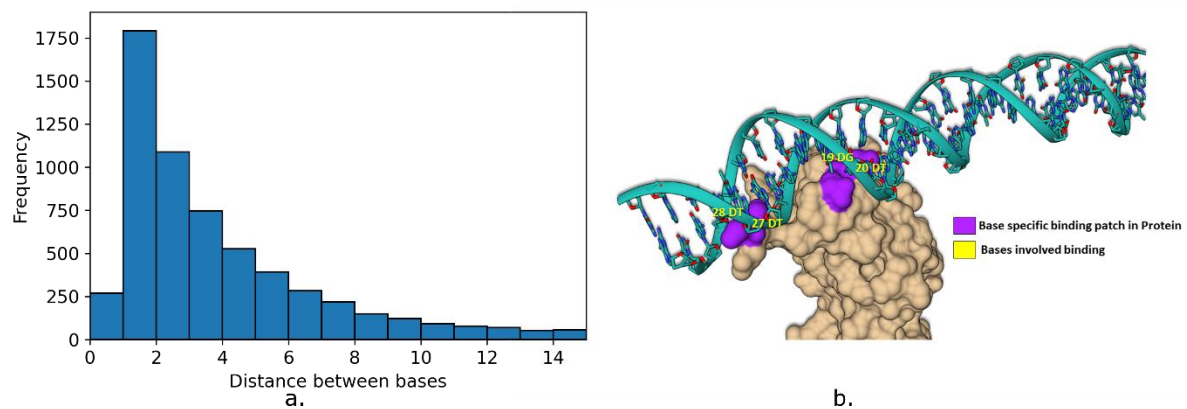


Figure 3.2: a. Frequency distribution of sequence distances between hydrogen-bonded nucleotides in an NR set of 285 DNA-bound protein complexes ². The x-axis is the distance between bases and the y-axis is the frequency. b. An example is where the distance between two DNA bases involved in a specific interaction is more than 6bps (PDB ID: 4jl3). The protein is shown in the surface representation and the DNA is shown in the ribbon and stick. The specific bases that make hydrogen bonds with the proteins are labelled (in yellow). Two distinct patches on the surface of the protein hydrogen bonded is shown in purple.

3.3.2 OE ratio highlights motif preference across chromosomes:

We had previously established that protein binding stretches in genomic DNA span 5-6 nucleotides² ([section 3.3.1](#)). So, in this study, we have read the genome at this length scale, i.e., at biologically relevant DNA motif sizes. To begin with, we compared the OE_{whole} values across motif sizes ranging from 2-6 bases. The motif size range was chosen to start from 2 bases to evaluate whether the observed distribution patterns are solely influenced by the base composition or if they are indicative of specific motif patterns ([figure 3.3 a-c](#)). We observed that with an increase in motif size, the number of motifs with high OE_{whole} values (> 2) along with an increase in the ranges of the

OE_{whole} (figure 3.4). From these observations, we infer that the variations in the OE_{whole} values are indicative of a more pronounced pattern of occurrence of motifs in size range 5-6. Hence, for the rest of the study, we focussed on 5 and 6 mer motifs and calculated the OE for different genomic ranges across all chromosomes.

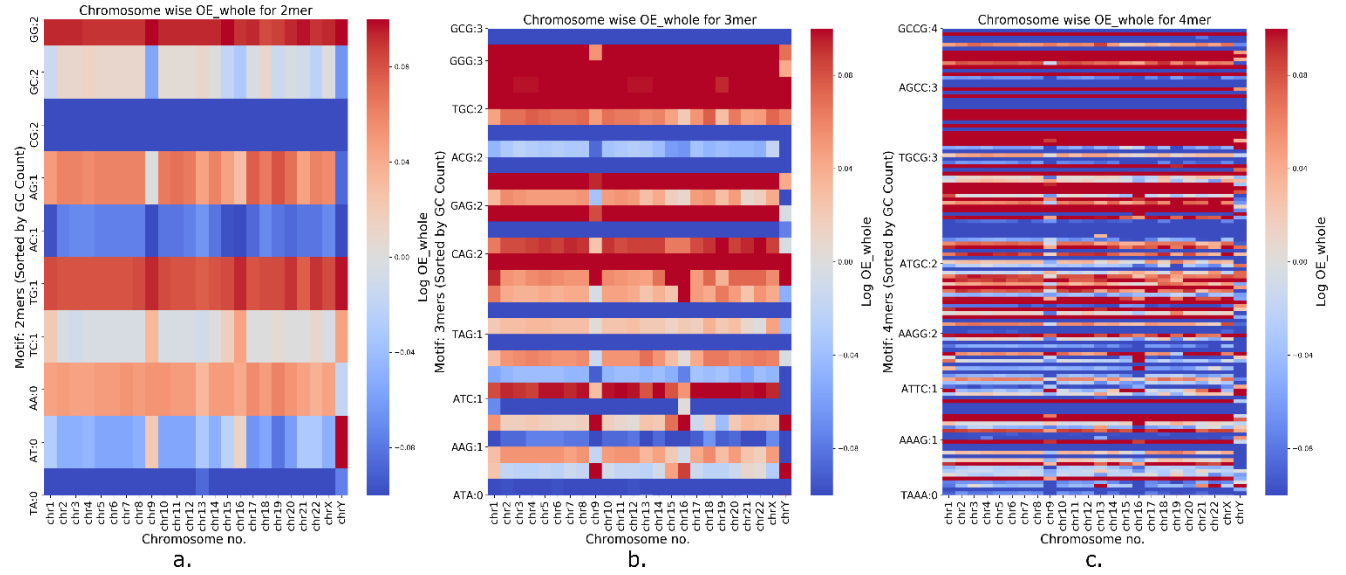


Figure 3.3: Heat maps of OE_{whole} ratios for motif size 2-4 (panels a, b and c respectively) across the different chromosomes shown in log scale. The log(OE) ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio. The motifs are arranged according to GC content, which increases going from bottom to top. In panels b and c, only a few representative motifs (and their reverse complements) are labelled. Shown alongside the motif labels are their GC contents.

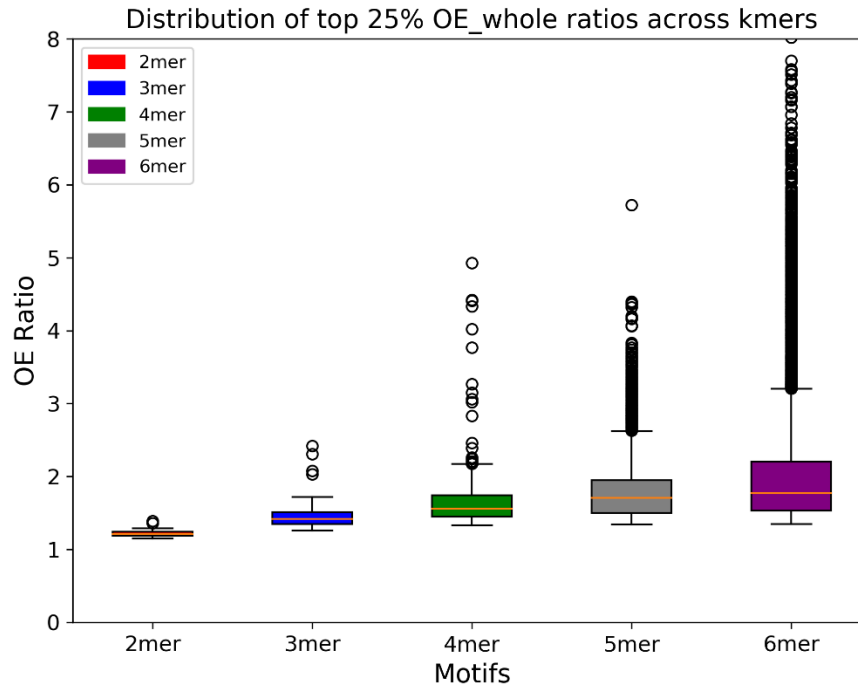


Figure 3.4: Box plot showing the distribution of the top 25% OE_{whole} for motif vectors of different sizes. The y-axis represents the OE_{whole} (limited to the range 0 to 10) and the x-axis represents different motif sizes. The different colours are used to represent different motif sizes.

The analysis of the OE_{whole} across different chromosomes highlighted a higher/lower abundance of common motif patterns that were independent of the ATGC content. The OE_{whole} motif vector consisted of the average values of a motif and that of its reverse complement as we are looking at double-stranded DNA. Hence, we got 512 and 2080 motif pairs for 5 and 6mers respectively. The number of 6mer motif pairs is not the exact half of all the 4096 motif combinations and the additional number is accounted for by the palindromic motifs where the forward and reverse complement pair are the same. On comparing the motif abundance across different chromosomes, we observe two trends. Firstly, certain motifs are more or less abundant independent of the chromosome of interest. The high or low abundance of the motifs can be inferred from their OE_{whole} ratios ([figure 3.5](#)). The 6mer motif CCCAGC consistently has a high OE_{whole} thereby a higher abundance across all chromosomes. The **CCCAGCAG** is a reported consensus site for zinc finger protein ZNF143⁴⁷. Secondly, we also observed that certain chromosomes have a variation in the trend of the OE_{whole} values. Chromosomes 9 and Y are the two such chromosomes which have distributions of motif OE_{whole} values starkly different from others ([figure 3.5](#)). This could be an attribute

of the difference in the structural variability and the occurrence of the largest heterochromatin composition in chromosome 9⁴⁸. The genes in chromosome 9 are reported to be scattered across highly repetitive sequence stretches^{49,50}. Likewise, chromosome 9 also harbours the maximum repeat region in their genomic sequence⁵¹. Thus, reading the genome in terms of motifs highlights the clustering of chromosomes and also the identification of motifs that may be key players in the recognition of different proteins⁵².

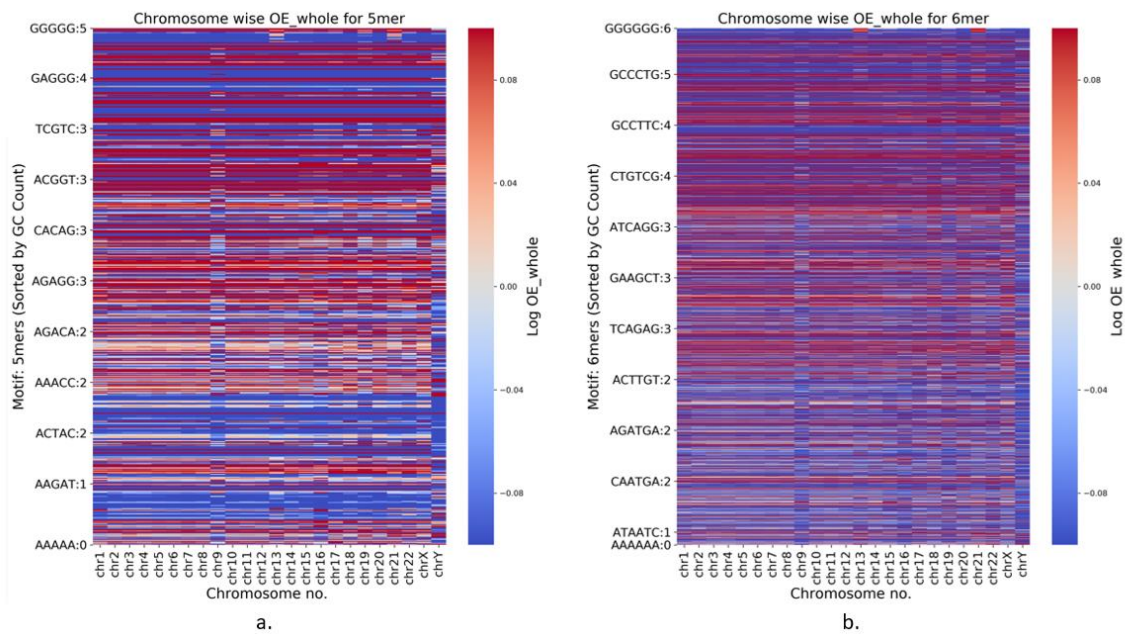


Figure 3.5: Heat maps of OE_{ratio} ratios for motif size 5-6 (panels a-b respectively) across the different chromosomes shown in log scale. The $log(OE)$ ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio. The 512 and 2080 motifs for 5-mers and 6 mers respectively are arranged according to GC content, which increases going from bottom to top. In panels, a and b, only a few representative motifs (and their reverse complements) are labelled. Shown alongside the motif labels are their GC contents.

3.3.3 OE ratio clusters chromosomes based on motif preferences:

We previously observed (in [section 3.3.2](#)) that certain motifs have high or low abundance in certain chromosomes based on their OE_{ratio} values for both 5 and 6 mers. We clustered the chromosomes based on the motif abundances. On clustering the chromosome, we find that the clustering pattern is similar for both 5 and 6mers with minor rearrangements. From the dendrogram representing the clustering in figure

[3.6a](#) and [3.6b](#), we observed that there is a large cluster of 14 chromosomes (Chromosomes 2-8,10-12,14,18, 20 and X) which are more closely related to one another than others considering the OE_{whole} . We also observed some smaller clusters of chromosomes which have similarities among themselves. Chromosomes (17,19), (1,15,16), and (13,21,22) are more similar to one another ([figure 3.6](#)). Chromosomes 9 and Y are the two most distinct ones and bear the least resemblance to the others (uniqueness of the chromosomes 9 and Y is already highlighted in [section 3.3.2](#)).

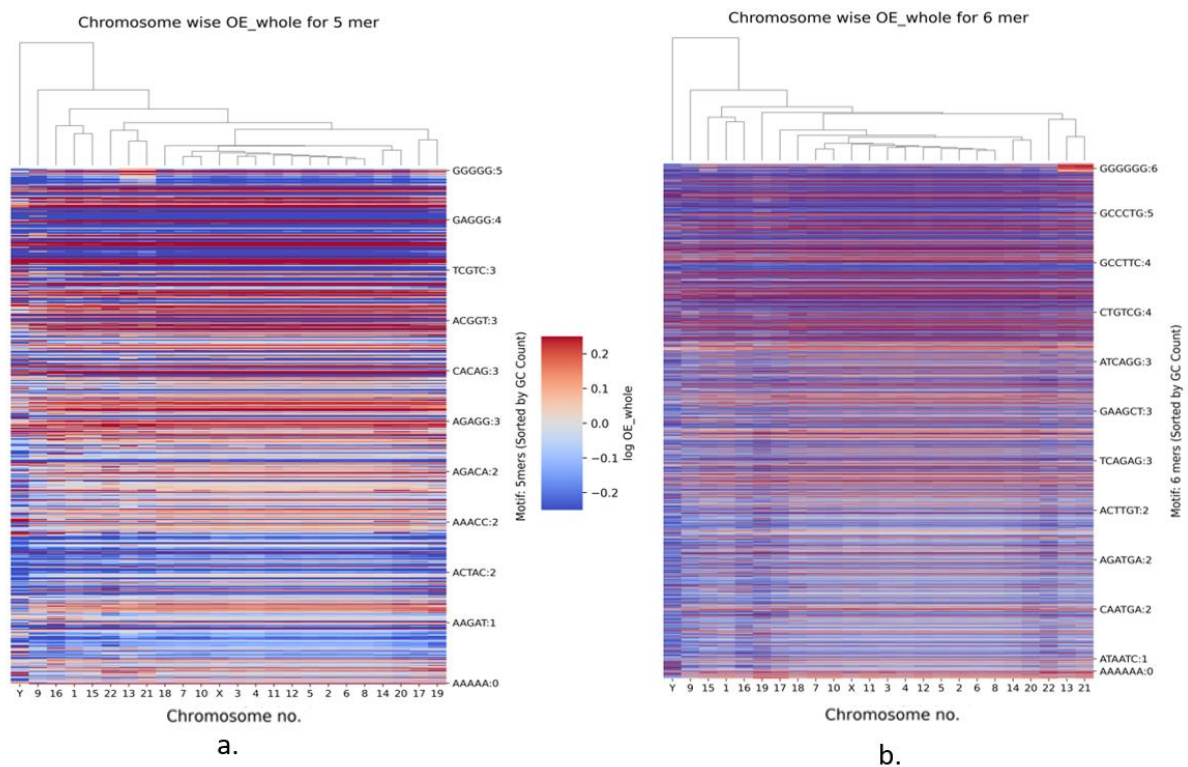


Figure 3.6: Heat maps of OE_{whole} ratios for motif size 5-6 (panels a-b respectively) across the different chromosomes shown in log scale with the clustering of chromosomes based on OE ratios. The dendrogram represents the clustering of chromosomes based on the similarity in motif distribution. The $\log(OE)$ ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio.

We also calculated the pairwise correlations between individual chromosomes to highlight the clustering pattern for motif sizes 5 and 6mers ([figures 3.6](#) and [3.7](#)). The correlation between chromosome pairs also highlights the grouping of chromosomes based on OE_{whole} , with some chromosomes more similar to one other compared to

others. The distinction between chromosome 9 and Y with other chromosomes is highlighted when compared with pairwise correlation as well.

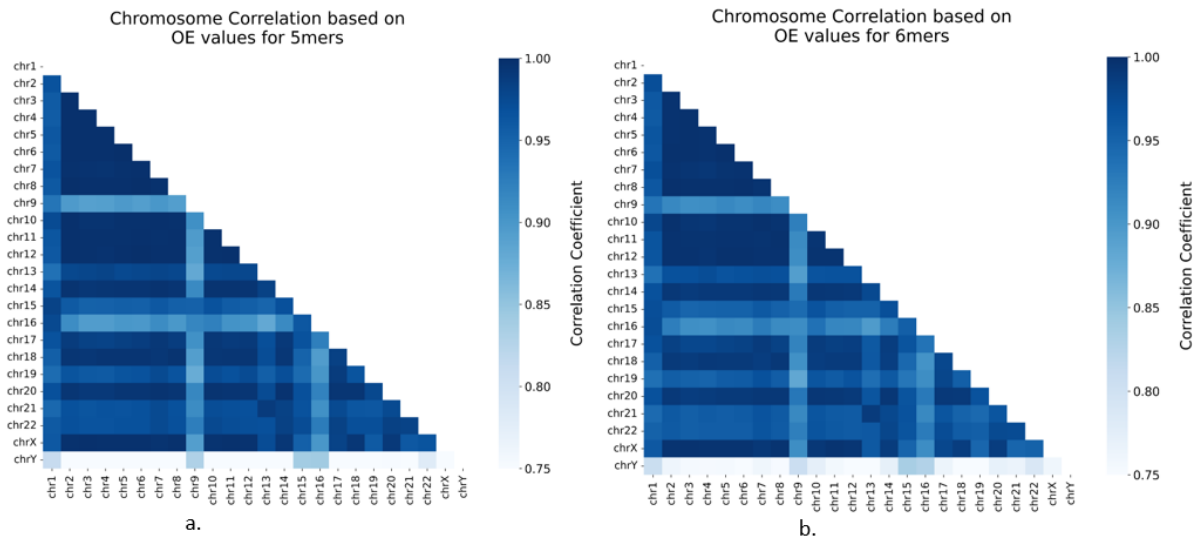


Figure 3.7: Heat maps of the Pearson's correlation coefficient (PCC) based on motif vectors in the whole chromosome for motif sizes 5 and 6 in 6a. and 6b. respectively. The x and y axes represent the chromosome pairs. The Pearson correlation coefficients are coloured white through to blue (colour legend). The darker the shade of the blue signifies a higher correlation.

3.3.4 Motif distribution in the genome is non-random:

From the results in sections [3.3.1](#), [3.3.2](#) and [3.3.3](#) it is evident that broad conservation patterns exist across chromosomes. We also observed that certain chromosomes have a unique motif distribution. From the whole genome, we now looked at the motif abundances at a resolution of 100kpbs stretches. We divided the chromosomes into ranges of 100kpbs bins and calculated the $OE_{100kpbs}$. Within each chromosome, sub-regions have different distributions of the motifs, as seen in the 6-mer motif distribution along 100kpbs stretches of chromosome 18 ([figure 3.8a](#)). Chromosome 18 has a size of around ~80.3Mbps. We retained the base composition of the original genome and generated a randomized genome by scrambling the original chromosome sequence. We calculated the OE_{random} based on the motif OE at 100kpbs resolution from the randomized genome. The distinct pattern in different 100kpbs stretches in the real distribution ($OE_{100kpbs}$) ([figure 3.8a](#)) was no longer visible for the randomized genome (OE_{random}) ([figure 3.8b](#)). We compared the randomization in all the chromosomes and

compared it with the original distribution of OE_{100kbp} (figure 3.9). The observation was consistent across all chromosomes. The only pattern that was consistent in the OE_{random} (figure 3.9) across all the chromosomes is the similarity of motif distribution based on the GC content and hence a gradient as the motifs were sorted in the order. The high resolution images in figure 3.9 can be accessed from (https://drive.google.com/drive/folders/1I3ajQtmGQTGDO1ZIRcm9XHqkcdMmYp7R?usp=drive_link)

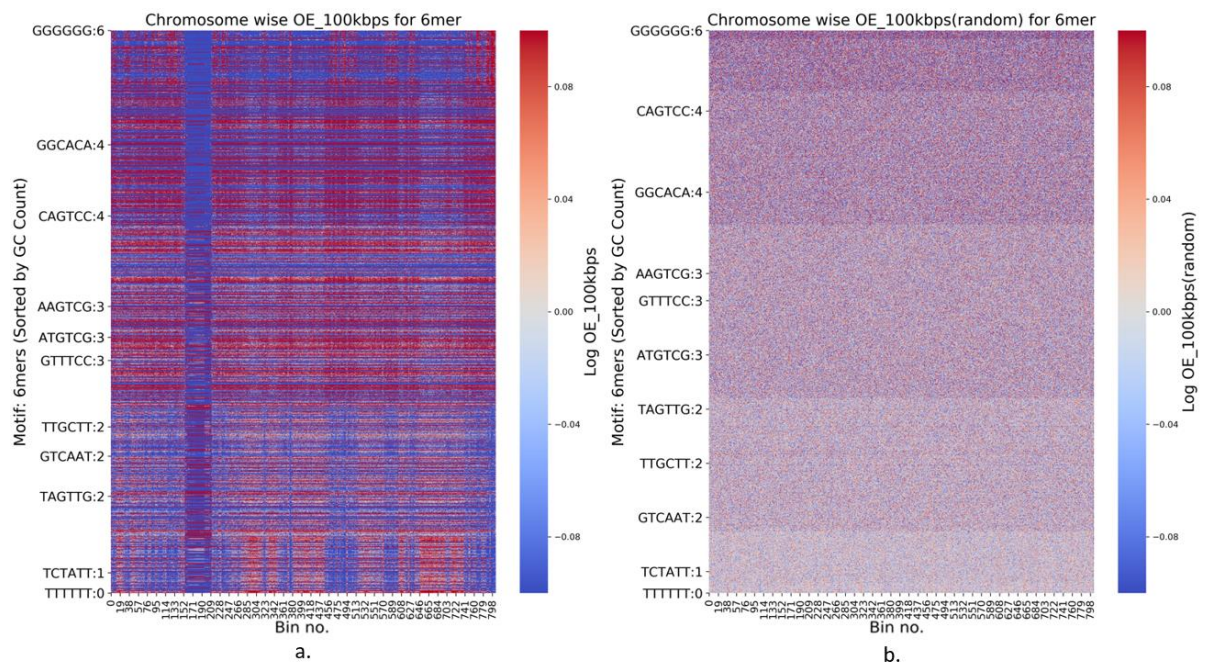
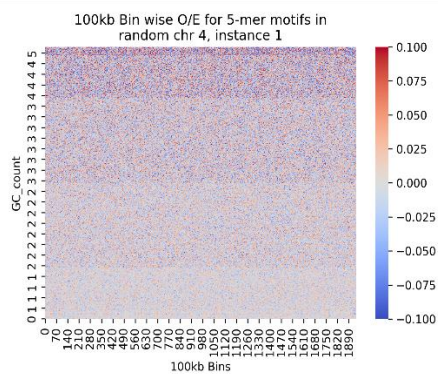
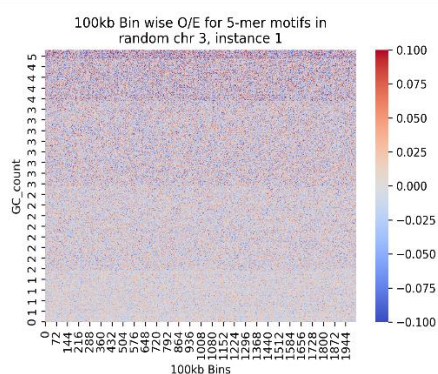
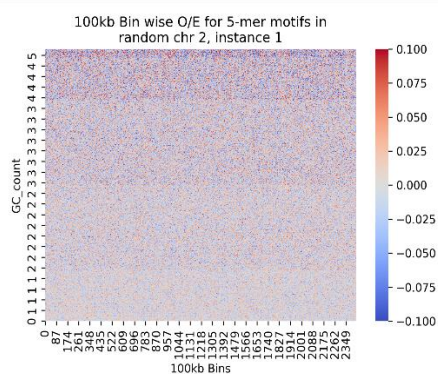
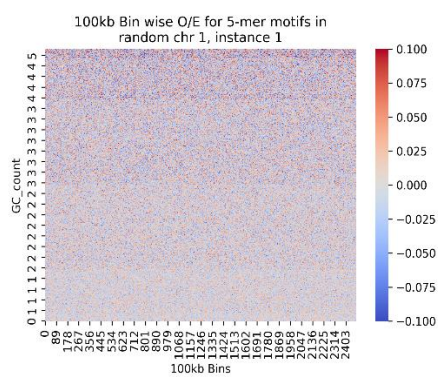
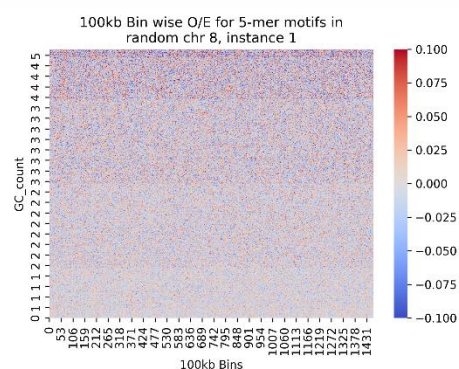
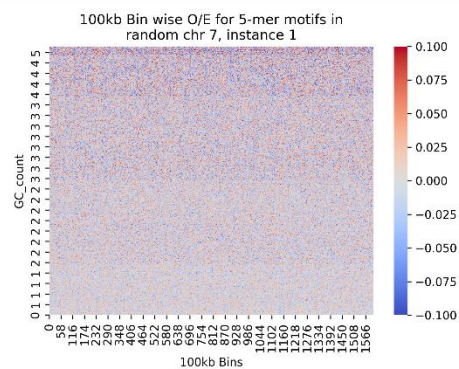
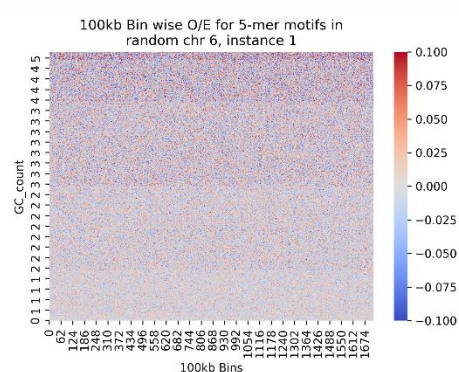
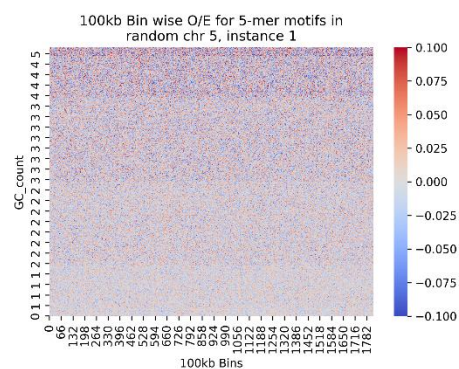


Figure 3.8: Heat maps of the log values of motif OE ratios. Panels a and b show the OE_{100kb} and OE_{random} heat maps for chromosome 18 in log scale respectively. The $\log(OE)$ ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio.

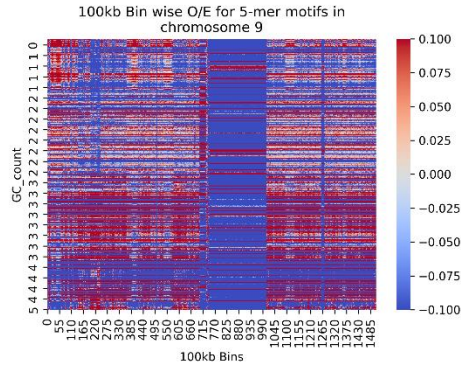
Random



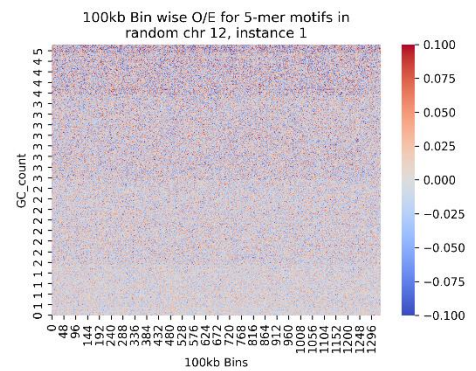
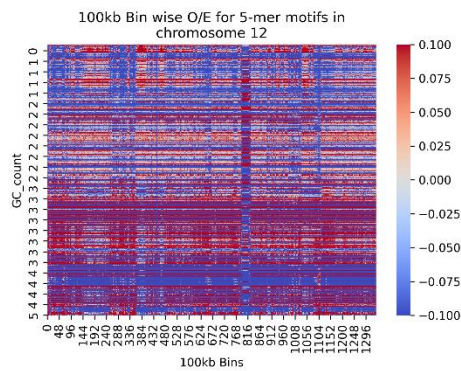
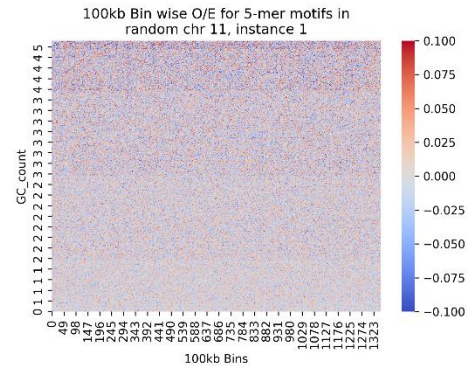
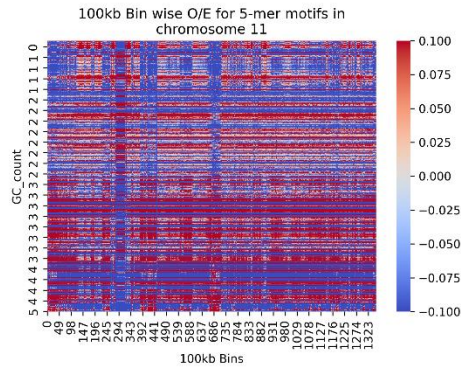
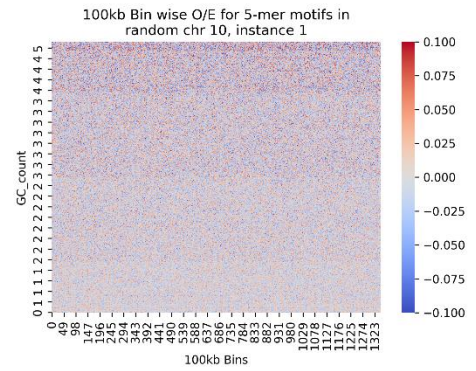
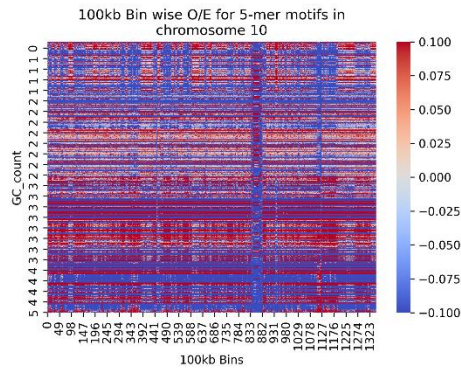
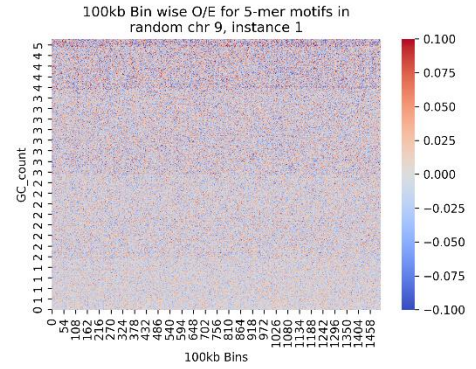
Random



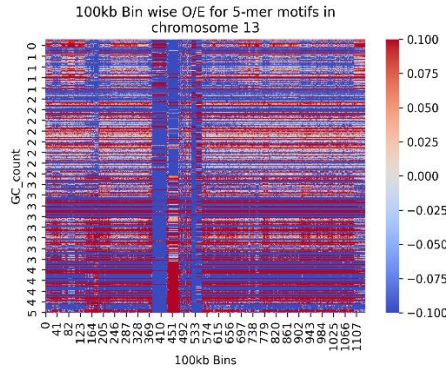
Original



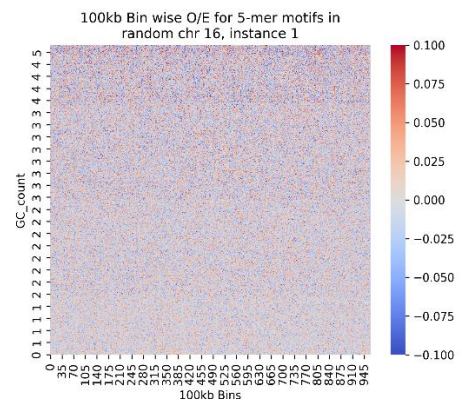
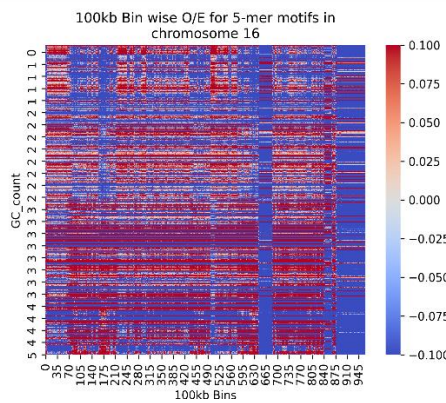
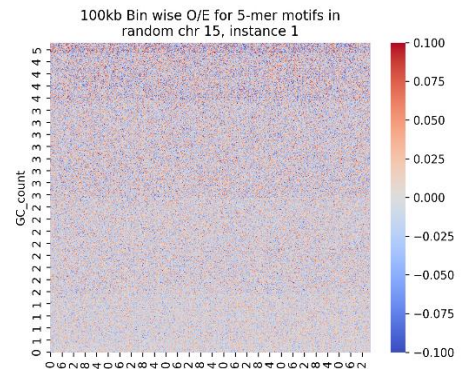
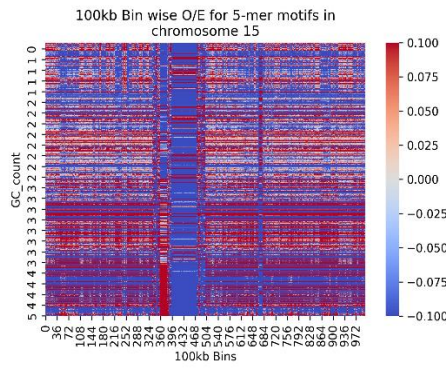
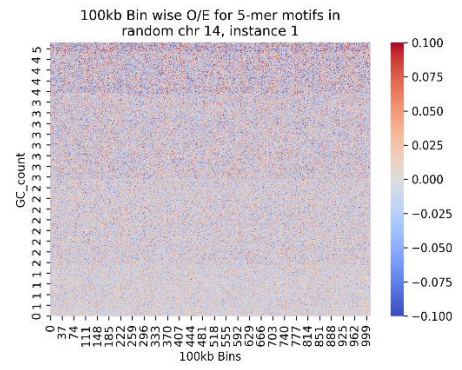
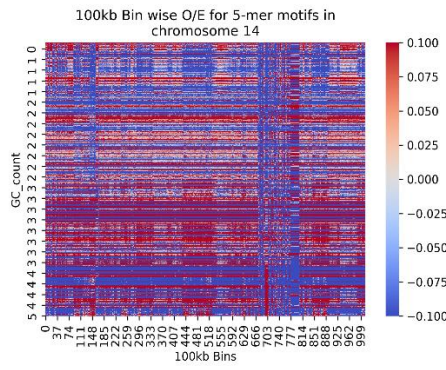
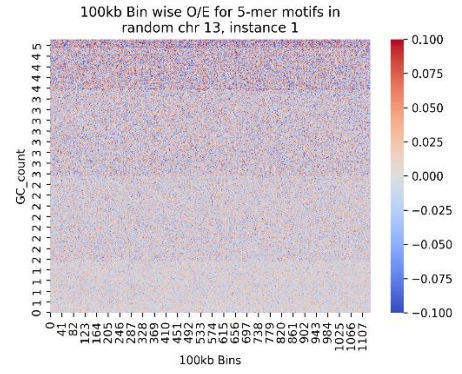
Random



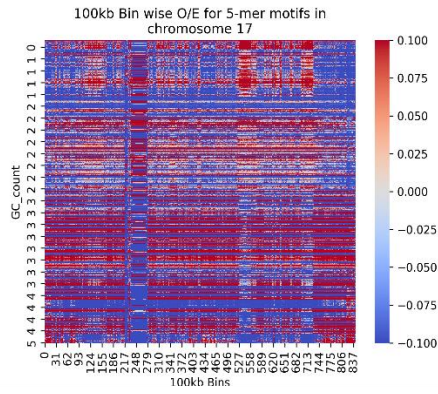
Original



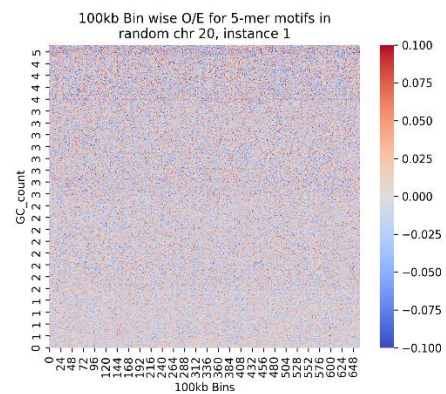
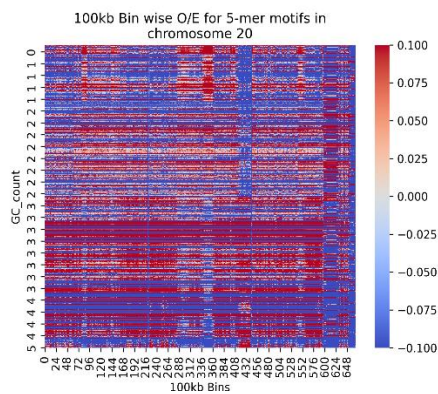
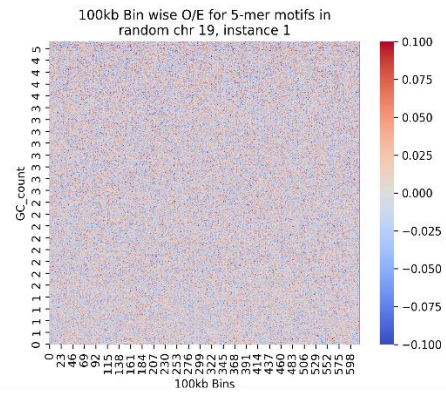
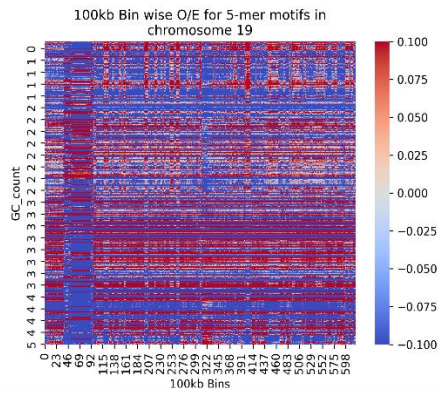
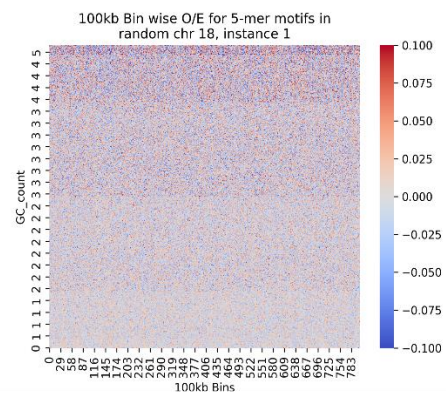
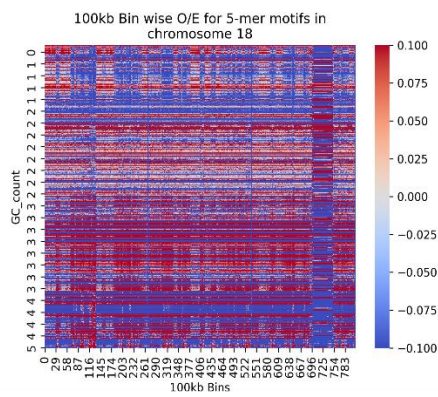
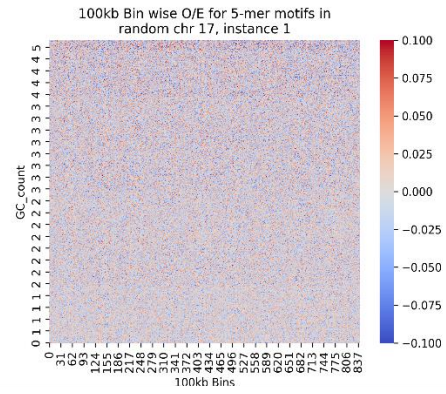
Random



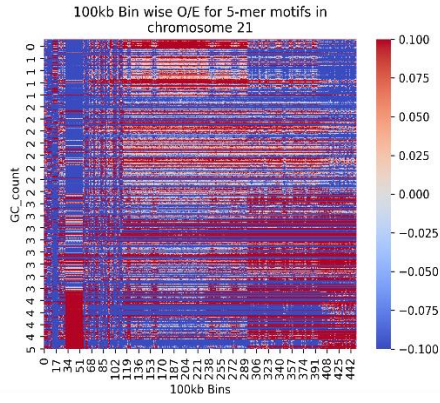
Original



Random



Original



Random

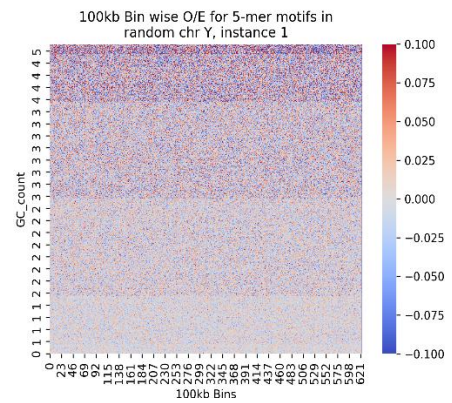
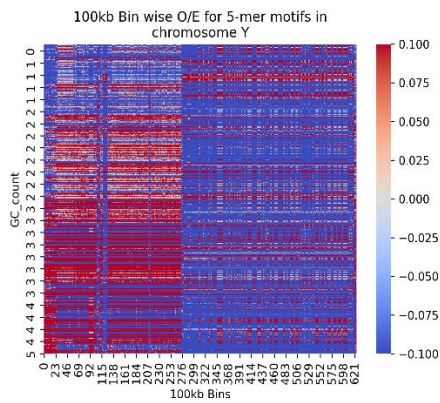
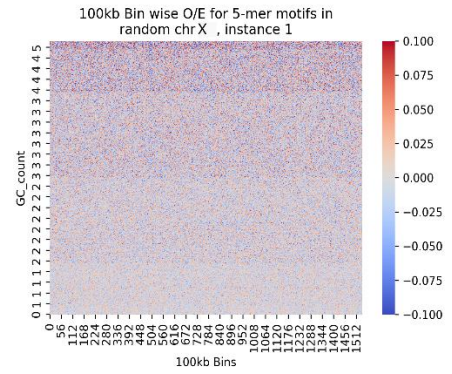
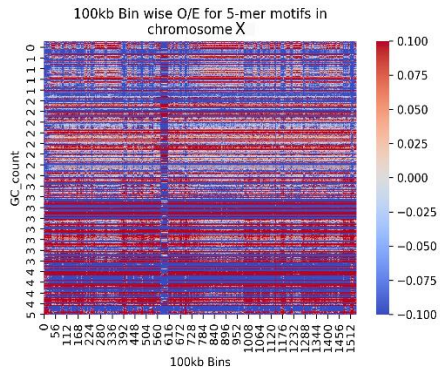
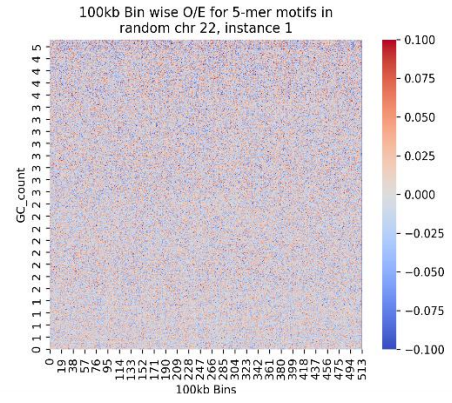
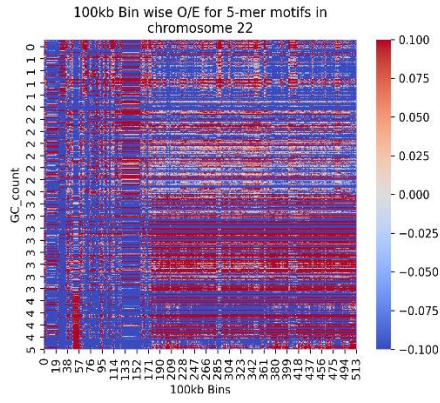
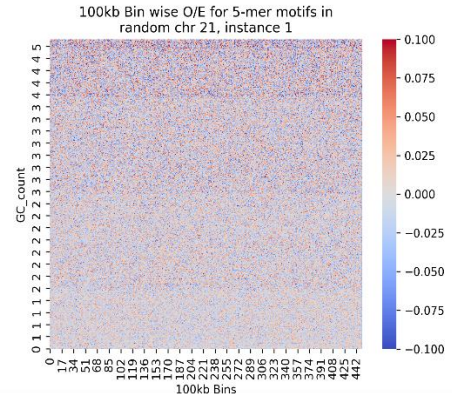


Figure 3.9: Heat maps of the $\log OE_{100kb}$ for all actual and random chromosome sequences, for motif size 5 mers. The panel on the left shows the plot for the actual chromosomes sequence and the panel on the right shows the results for randomised chromosome sequence. The $\log(OE)$ ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio. The 2512 motifs for 6-mers are arranged according to GC content, which increases going from bottom to top. On the y-axis, the GC content of the motifs is labelled.

From the results above it is clear that the variation in motif distribution is also evident within chromosomes. We next investigated the centromeric regions across all the chromosomes. We obtained the centromeric boundaries from the NCBI genome data viewer and UCSC genome browser^{46,53}([table 3.1](#)). We observed that the $OE_{centromere}$ distributions across the different chromosomes of the genome are different ([figure 3.10a](#)). We repeated the same analysis on the centromeric region from the randomised chromosome (retaining the similar base composition as the original chromosome) ([figure 3.10 a, b and c](#)). We observe that the distinct patterns visible in the real distribution of $OE_{centromere}$ are no longer apparent in the randomized case (OE_{random}). We repeated the randomization in triplicates and the observations were consistent ([figure 3.10](#)). We also calculated the pairwise correlations between individual chromosomes to highlight the clustering pattern for motif sizes 5 and 6mers (figures 3.11a and 3.11b). The correlation between chromosome pairs also highlights the grouping of chromosomes based on $OE_{centromere}$. For instance, the acrocentric chromosomes 13,14,15,21 and 22 have a distinctly different $OE_{centromere}$ pattern ([figure 3.10a, 3.11a and 3.11b](#)). It is known that the short arms of these acrocentric chromosomes are abundant with segmental duplications²⁰, which could be the explanation for our observation.

Table 3.1: Centromere boundaries for $OE_{centromere}$ calculation. Column 1 represents the gene name. Columns 2 and 3 represent the start-stop boundaries for each chromosome respectively. Columns 4 and 5 show the chromosome-wise centromere size in bps and Mbps respectively.

Chromosome	Start	Stop	Size(bps)	Size/Mbps
chr1	121973252	125421848	3448596	3.45
chr2	91879019	96296965	4417946	4.42
chr3	87983662	93990946	6007284	6.01
chr4	48247029	51956625	3709596	3.71
chr5	46259962	51541276	5281314	5.28
chr6	58631981	62603545	3971564	3.97
chr7	58365242	62251569	3886327	3.89
chr8	43210183	47354764	4144581	4.14
chr9	42215186	45645422	3430236	3.43
chr10	38050203	41649353	3599150	3.6
chr11	51079005	55858265	4779260	4.78
chr12	33334329	37896869	4562540	4.56
chr13	16502400	18949340	2446940	2.45
chr14	16100180	18266678	2166498	2.17
chr15	17458137	20558059	3099922	3.1
chr16	35322745	38439712	3116967	3.12
chr17	22694097	27672900	4978803	4.98
chr18	15431301	21560433	6129132	6.13
chr19	24221370	28153079	3931709	3.93
chr20	25773360	30501863	4728503	4.73
chr21	10901609	13034530	2132921	2.13
chr22	13679758	17432904	3753146	3.75
chrX	58278961	63921024	5642063	5.64
chrY	10300948	10664448	363500	0.36

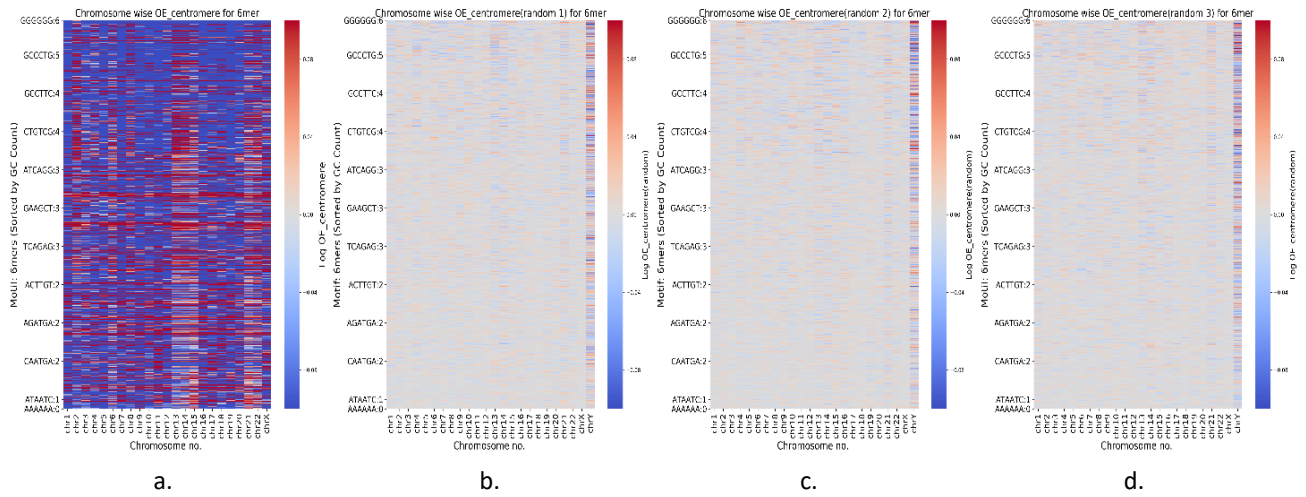


Figure 3.10: Heatmaps for motif distribution based on $OE_{centromere}$ across different chromosomes. Panel a shows the motif distribution in the original sequence. The 3 sets of randomised sequences of the same centromeres are shown in d-f. The $\log(OE)$ ratios are coloured red through to blue (colour legend) where red and blue indicate values of > 1 and < 1 respectively. The darker the shade of the red (or blue) the higher (or lower) the ratio.

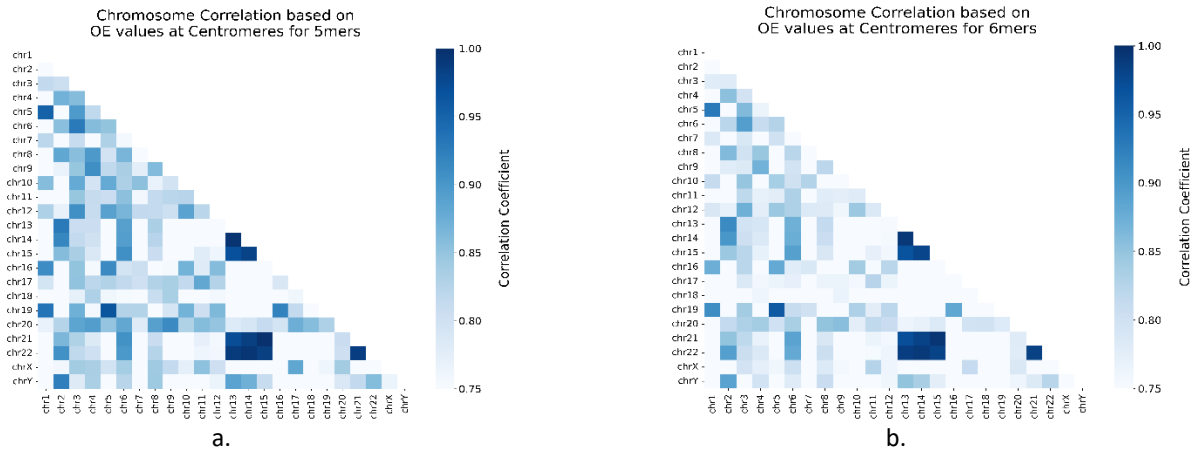


Figure 3.11: Heat maps of the Pearson's correlation coefficient (PCC) based on motif vectors in the centromeres for motif sizes 5 and 6 in a. and b. respectively. The x- and y-axis represent the chromosome pairs. The Pearson correlation coefficients are coloured white through to blue (colour legend). The darker the shade of the blue signifies a higher correlation.

3.4 Discussion:

We have attempted to understand the functional significance behind the genome sequence. We studied parts of the genome just as they would be recognised by the

interacting protein partners, i.e., 5 to 6 base pairs at a time. We established this optimal recognition size by studying 285 NR DNA-protein complexes. Thus, DNA binding proteins recognize their cognate sites on genomes by making interaction with either 5 or 6 bases per domain. There are exceptions in the size of consensus but these are usually in cases where the DNA binding domains dimerise.

The findings from our comprehensive analysis of motif distributions across and within chromosomes shed light on the regulatory landscape of the genome and reveal potential coordination among different chromosomes. We studied the distribution pattern of a total of 512 and 2080 unique 5-mers and 6-mers respectively that were obtained by averaging the values of motifs and their corresponding reverse complements. We observe a clustering of chromosomes based on this distribution of motifs. This clustering was based on the two observations. Certain motifs are found occurring in high or low abundance across all chromosomes. We also observe that within a chromosome there is a distinct pattern of motif distribution.

We have shown that the distribution of these 5/6-mers is non-random. In real genomes, there are regions of the genome where certain motifs are over or under-represented. We have quantified this using the OE ratios. The non-randomness of the motif distributions happens at the scale of the whole chromosome or even at the level of smaller segments of the genome such as 100kbps stretches or even around centromeric regions across all chromosomes. Using the simple metric of $OE_{\text{centromere}}$ ratios, we establish a unique pattern of motif distributions conserved across the different acrocentric chromosomes (13,14,15,21 and 22). The centromeres of these chromosomes are known to have segmental duplications and our analysis using only the motif distributions in these regions also identifies the similarities. This study highlights that distinct motif distributions ensure differential binding by the different protein partners.

Acknowledgement

The analysis of genome randomization and the dendrogram of chromosome clustering was done by Sumant Chopde.

Chapter 4

Promoter correlation to gene regulation

- [Gene correlations from promoter motif distributions](#)
- [Significance of gene promoter correlations](#)
- [Features of gene promoters having high correlations](#)
- [Gene pairs with high correlation share common transcription factors](#)
- [Multiple gene regulation by a common transcription factor](#)
- [Promoter correlations highlight spatial proximity](#)

4.1 Introduction:

The current T2T-CHM13 genome assembly has annotated a total of 63,494 genes within the human genome. This count encompasses both protein-coding genes, pseudogenes, hypothetical genes and functional RNA (lncRNA, miRNA etc.) genes and genes in which predicted frameshift mutations have been considered¹¹. The transcriptional regulation of the different biological processes is a result of a dynamic process of initiation, elongation and termination along with chromatin accessibility and epigenetic mechanisms⁵⁴. The promoters are DNA regulatory regions located upstream of a gene transcription start site, which contributes as control points for gene transcriptional regulation⁵⁵. In this study, we extended our analysis by looking at DNA 5/6-mer motif abundances in the gene promoter region based on the established patterns in the genome.

The regulation of the transcriptome inside the cell nucleus is complex²². To ensure successful gene transcription, it involves the coordination of a complex network involving the transcription factors, chromatin regulators, and enhancer-promoter looping^{1,56}. There are examples of spatial proximity of genes, which are distant sequentially, as a result of coordination by different protein interactions⁵⁷. Thus, we extend our motif distribution analysis to compare genome segments in the gene promoter to get insight into the gene network. We compare the motif abundances in the promoter regions of all the 63,494 annotated genes to find gene promoters that share repetitive motifs, thereby establishing a possibility of being recognized by a protein binding partner. Thus, our calculations establish a network of 63,494 genes based on the motif observed expected (OE ratios) ([figure 4.1](#)). We use correlation scores to signify the level of similarity between two gene promoters.

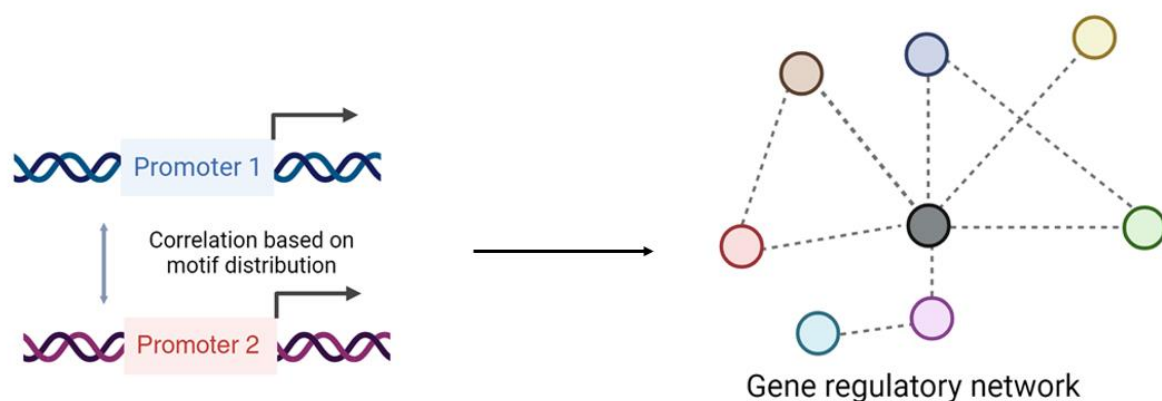


Figure 4.1: A schematic of the using motif vector distributions in gene promoter to generate a gene regulatory network of all annotated genes in the human genome.

We validated our findings using three case studies. In the first case study, we selected three genes, LPHN1(ADGRL1), CDK9 and TRIM8, that exhibited high correlation based on our analysis using motif abundances and compared them to a network map using NetworkAnalyst^{58,59}. We identified two common transcription factors to all three genes, thus validating our method and its use to find functionally significant gene connections. For the second case study, we used a larger set of 19 genes which are reported to be regulated by the same transcription factor(s) Jun/Fos. Our analysis highlighted the correlation between the gene clusters. Thus, the method can be used to predict differential gene regulation. The third case study investigated the use of the method to predict the spatial proximity of genes. We used Hi-C datasets which provide insights into genome compartmentalization⁶⁰. We validated our correlation with Hi-C along with promoter capture Hi-C data (pcHi-C) datasets. Promoter capture Hi-C which captures distal promoter interacting regions for all promoters⁵. We looked at the correlation of genes that overlap with these regions detected by the Hi-C experiments. We found a strong correlation between our scoring and the spatial positioning, reinforcing the significance of our methods.

Utilizing the updated annotations of human genome sequence in our analysis provided a clearer understanding of the functional significance of motif localization concerning genes and how it influences cellular organization and biological processes. Overall, our study sheds light on the intricate relationships between DNA motifs, gene regulation, and genome organization. The unbiased approach, validation through

multiple case studies, and utilization of the updated genome sequence contribute to the robustness and reliability of our findings.

4.2 Methodology:

4.2.1 Motif distribution in gene promoters:

We defined the promoter regions into three categories based on size: 1kbps, 2kbps and 6kbps upstream the gene start site. We used the OE ratio which is a statistical method used to measure the extent to which a particular feature is overrepresented or underrepresented within a given dataset. The OE ratio is calculated as:

Observed count (O): Number of times a motif occurs in a genomic stretch

$$\text{Expected count (E)} = (N - k + 1) \prod_x f_x^{n_x}$$

$$\text{OE Ratio} = \frac{\text{Observed count}}{\text{Expected count}}$$

Where:

n_x : number of times a base occurs in a motif

f_x : base probability in the genome

k : motif length (in base pairs)

N : chromosome length (in base pairs)

For the different promoter sizes, we obtained OE_{1kbps_kmer}, OE_{2kbps_kmer} and OE_{6kbps_kmer} for gene promoter sizes 1, 2 and 6bps and for motif size $k = 5$ and 6 respectively. We obtained the coordinates of gene start and sites from the NCBI RefSeq gene annotation^{46,53} and the promoter regions were extracted accordingly.

4.2.2 Comparing motif vectors of gene promoters:

To conduct a comparative analysis of motif patterns within different gene promoters of the same chromosome and across different chromosomes, we obtained the OE_{promoter} vector list of size k-mer count. To identify any trends shared among a set of promoters within the same chromosome or across different chromosomes, we calculated a correlation of the vectors. We compiled a dataset of promoter correlations for each k-mer vector of OE_{prom_1kbps}, OE_{prom_2kbps} and OE_{prom_6kbps} for motif sizes 5 and 6. The final

dataset comprised $N_i \times 24$ chromosomes where N_i represents the number of genes on each chromosome. Correlations between motif vectors of different promoters were computed as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y} \text{ where } cov(X,Y) = \sum (X - \mu_X) \times (Y - \mu_Y)$$

Where $cov(X,Y)$ is the covariance, μ_X and μ_Y is the mean of OE of promoter X and Y and σ_X and σ_Y are the standard deviations of motif vectors of promoter X and Y respectively. This formula allowed us to quantify the degree of motif pattern similarity between different promoters and determine their correlation. By applying this methodology, we aimed to elucidate the presence of shared patterns and regulatory relationships among gene promoters across different chromosomes.

4.2.3 Finding Hi-C contacts with genes:

Hi-C data were obtained from GEO accession number GSE18215⁶⁰ (in .mcool format), where chromosomes were divided into uniform bins of different sizes. We chose to use data from the bins of the smallest size (highest resolution), 10kbps. We extracted pairs of 10kbps bins from the data where cross-links were recorded between chromosomes 18 and 19.

For the promoter capture Hi-C, we extracted the chromosome contact start-stop coordinates from the .bedpe files from GEO accession number: GSM1704495⁶¹. From the NCBI gene annotation for the corresponding reference assembly used in mapping the contacts, we obtained the gene coordinates that overlap with the contact regions between the two chromosomes. For the gene pairs that fall in the region of contacts, we obtained the promoter correlations according to our motif vector calculations at the different aforementioned resolutions.

4.3 Results:

4.3.1 Gene correlations from promoter motif distributions:

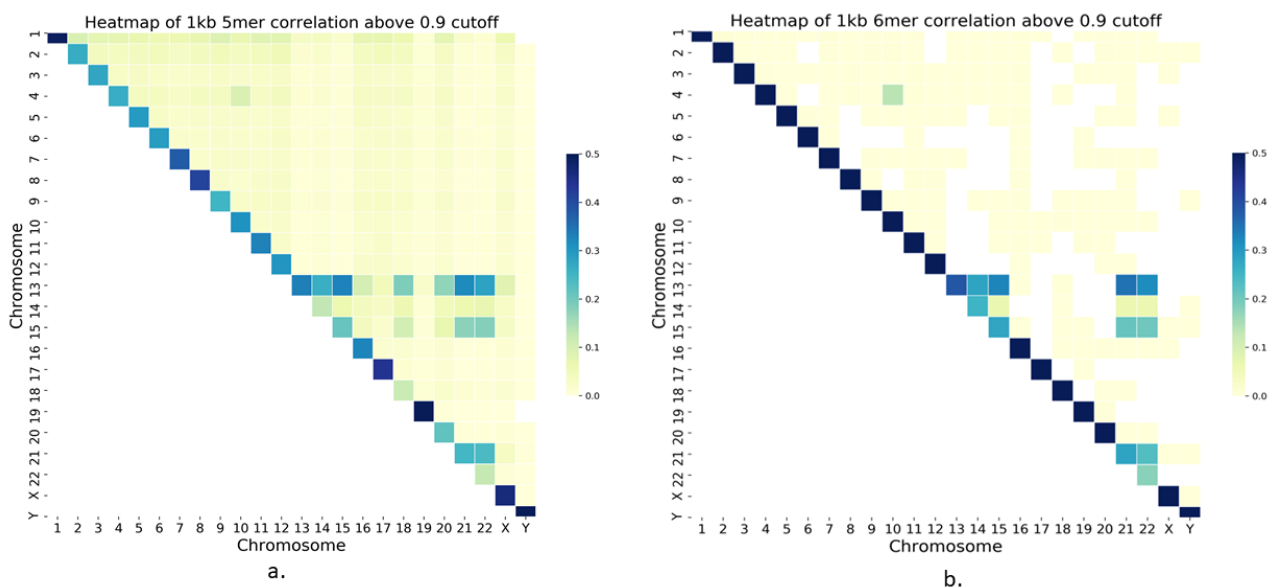
We defined the promoter regions as the regions 1kbps, 2kbps and 6kbps upstream of a gene start site. The size ranges were chosen as 1kbps, 6kbps (a maximum known effective size⁶²) and a value in between 2kbps. We obtained the abundance of the

motif vectors using the 2080 $OE_{Xkbps_k\text{-mer}}$ values for k-mers (k= 5 and 6) respectively, where X is the size of the upstream regions considered as promoters. We calculated the correlation between genes and the results were obtained chromosome-wise. Of all the ~4 billion gene-gene motif vector correlations, we observed that about ~25% of the motif vector correlations have a correlation coefficient of ≥ 0.9 (for both 5 and 6mers) ([figure 4.2](#)). The order of magnitude of these high correlations (≥ 0.9) is consistent for all motif sizes at different resolutions. As the promoter size increases from 1kbps to 6kbps the proportion of the high correlations between chromosome pairs changes but the trend remains the same ([figure 4.2 a-f](#)). A major contributor to the high correlations comes from gene promoter motif vector correlations from the same chromosomes. This is an attribute of the fact that most of the functionally co-regulated genes are part of the same chromosome and the localization of the genes is also proximal to one another. The average number of gene pair motif vector correlations above 0.9 is ~920 genes per chromosome pair. Thus, on average 920 unique gene pairs share correlations based on motif abundances in their promoter regions higher than 0.9.

We observe certain trends between chromosome pairs that contribute to a high proportion of the high correlations. The maximum number of inter-chromosomal gene pair correlations is 36,408 gene pairs between chromosome 13 and 21. On the other hand between chromosome Y and chromosomes 1 and 19, no gene pairs correlate with a score of 0.9 and above. From the inter-chromosome interactions, it is important to study the genes from different chromosomes that show similarity in the abundance of the motifs from the motif vectors. This implies that the genes could be co-regulated i.e. recognized by the same interacting partner.

We observe that the acrocentric chromosomes (13,14,15,21 and 22) share a large proportion of genes, which share a high motif vector correlation (≥ 0.9). There are 1768, 2216, 2072, 992 and 1310 genes in chromosomes 13, 14, 15, 21 and 22 respectively, that are correlated to at least two promoters from genes of other acrocentric chromosomes. This observation can be suggestive of the Robertsonian translocations^{63,64}, where the long arms of the chromosomes are joined to one another and which is unique to the chromosome group. Ribosomal RNA genes constitute ~4-13% ([table 4.1a](#)) of the total genes in chromosomes 13, 14, 15, 21 and 22. Additionally, we also checked what fraction of the total high correlations (Pearson's correlation

coefficient ≥ 0.9) were contributed by the rRNA gene pairs from both chromosomes. We observed that ~92-92% of the high correlations are contributed by the rRNA genes from both the chromosomes for chromosomes 13,14,15 and 21([table 4.1b](#)). However, none of the rRNA genes from chromosomes 13,14,15 and 21 with the rRNA genes of chromosome 22 contribute to the high correlations (table 4.1b). Thus such tandem repeats of rRNA genes highlight the conserved patterns. Similarly, chromosomes 4 and 10 also show a higher proportion of gene pairs (1017) which had a correlation coefficient above 0.9 ([figure 4.2a](#)). A translocation between these two chromosomes is also known in certain types of leukaemia^{65,66} and Wolf-Hirschhorn syndrome⁶⁷. However, the high correlations between the genes in these chromosomes are not an attribute of the rRNA genes. Just as in acrocentric chromosomes, the reason for the translocation could be the local similarity in the motif distribution.



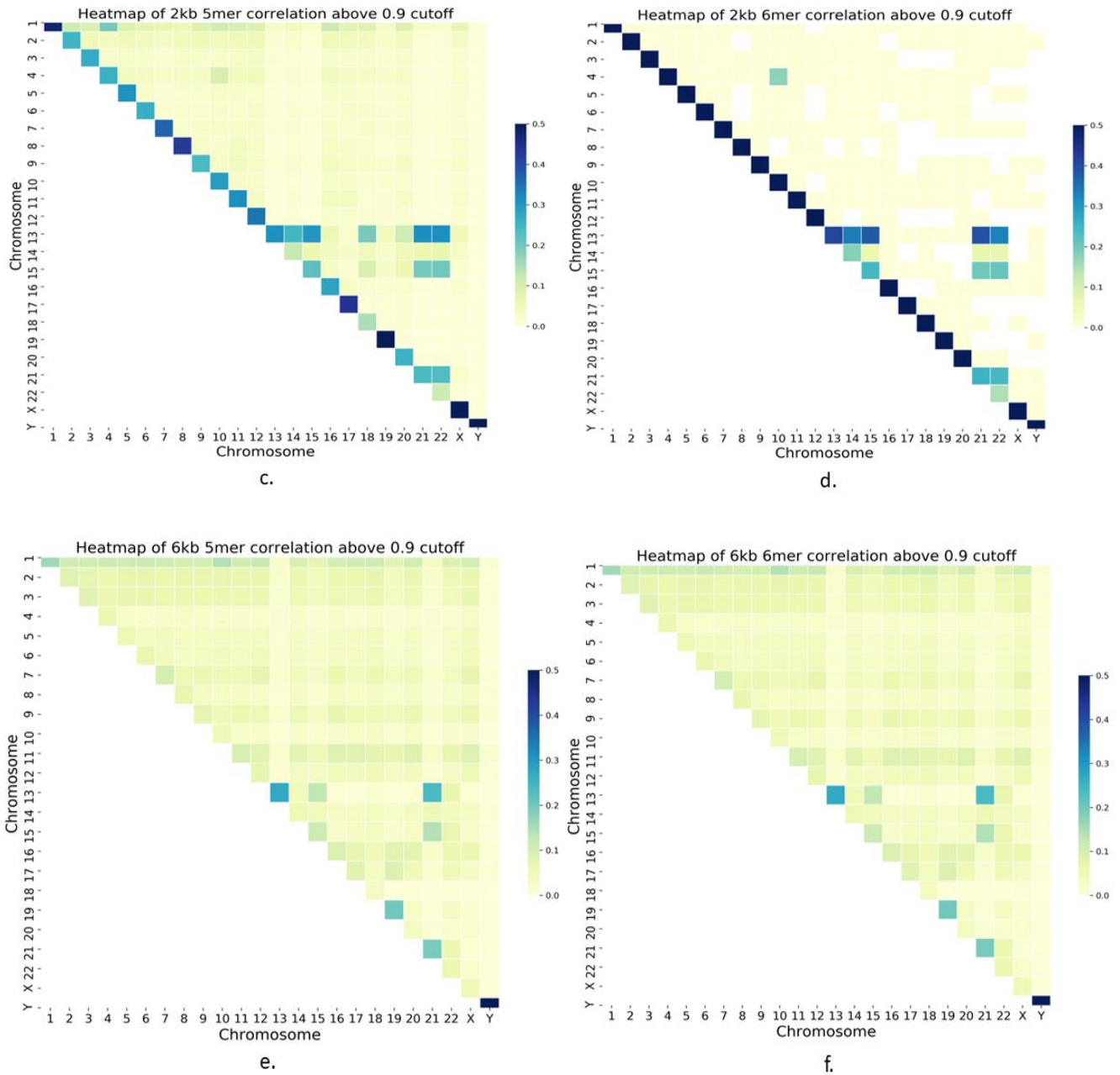


Figure 4.2: An inter-chromosome motif vector correlation map of frequency of gene pairs with correlation coefficients > 0.9 for k-mers with promoter sizes of n kbps (where n = 1, 2 and k = 5 and 6). The proportion of highly correlated genes are colour-coded on a white-yellow-green-blue gradient. The white end represents no correlation while the darker the shade of blue the larger the number of correlated gene pairs. Figure 1a-f shows the result for 1, 2 and 6 kbps for 5 and 6 mer motif sizes respectively.

Table 4.1a: Chromosome-wise composition of rRNA

Chromosome	rRNA genes
13	12.8%
14	12.2%
15	7.1%
21	15.4%
22	4.4%

Table 4.1b: Proportion of rRNA genes contributing to high correlations (≥ 0.9)

Chromosome	14	15	21	22
13	82.70%	91.50%	83.7%	87.4%
14		81.50%	81.5%	80.9%
15			86.10%	87.5%
21				82.3%

4.3.2 Significance of gene promoter correlations:

We wanted to study the feature of motif abundances in the regions proximal to the gene start and stop site. Hence, as a control, we repeated the analysis of motif vector correlations beyond the promoter regions. We repeated the analysis for downstream regions of genes and random bins from the chromosome of the study. The dimension of the random correlations will be the same as the number of genes in the chromosomes selected. For this analysis, we selected the chromosome pairs 6-7, 12-13 and 11-18. The selection of the chromosome pairs was done such that the sample set had chromosomes of different sizes and gene density ([figure 4.3](#)). We obtained the correlations at different resolutions of motif vectors of 5-mer and 6-mer for 1kbps, 2kbps and 6kbps regions in gene promoters, downstream regions and randomly selected regions. It is to be noted that the random regions may overlap with the genic regions as we have not eliminated any genic bins in this analysis. The higher correlation coefficients are more evident in the gene promoter regions as opposed to

the random regions or the downstream regions of the genes ([Table 4.2a](#)). It is to be noted that the gene pairs showing high correlations in the downstream regions can be an attribute of those genes where the start site of one gene overlaps with the stop site of another gene ([Table 4.2b](#)). In the case of a random bin set, the enrichment is consistent across all randomization and all resolutions.

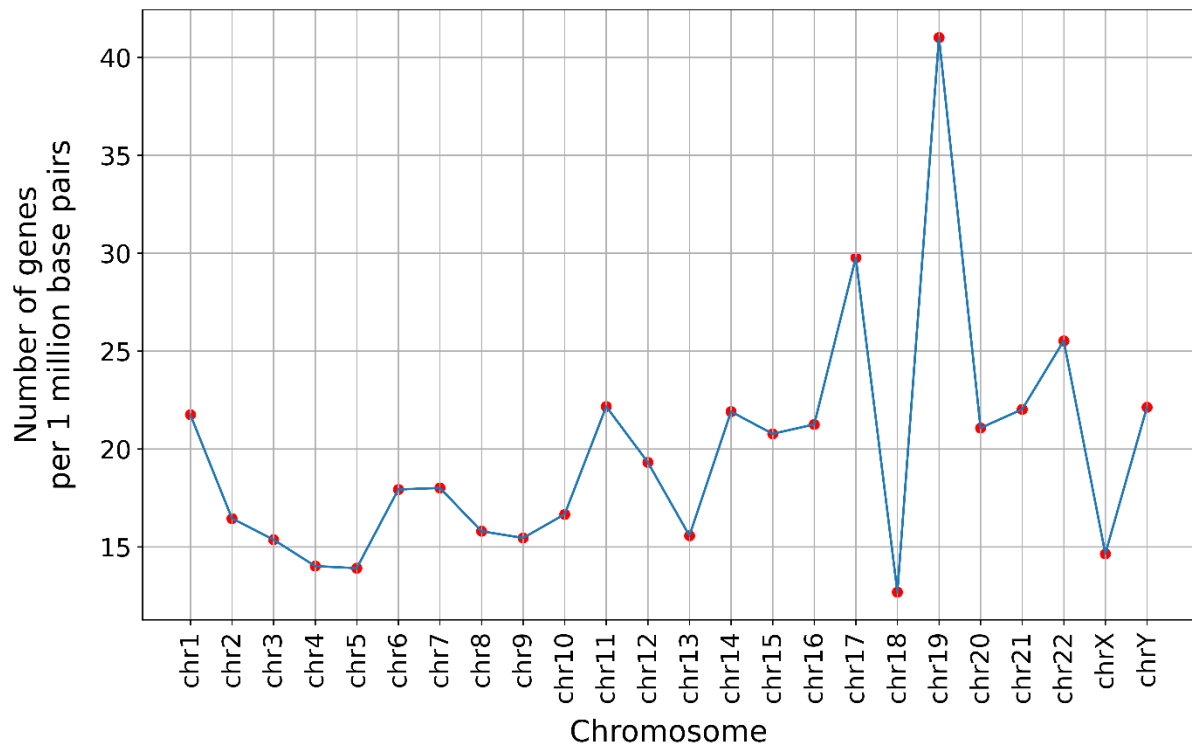


Figure 4.3: Chromosome-wise gene density plot. The y-axis represents the number of genes per Mb of the chromosome. The x-axis represents the different chromosomes. The data points represent the number of genes/Mb (gene density) of each chromosome.

Table 4.2a: Comparison of gene-pair correlation above 0.5 in promoter and downstream vs random bins.

Promoter Size	Motif Size	Genomic Region	Chr6-Chr7	Chr12-Chr13	Chr11-Chr18
1kb	5	Promoter	11.74	11.10	11.01
		Downstream	11.00	9.05	10.04
		Random	5.45	4.28	5.33
	6	Promoter	0.98	1.97	0.99
		Downstream	0.59	0.40	0.41
		Random	0.23	0.18	0.15
2kb	5	Promoter	33.63	29.66	32.60
		Downstream	39.23	32.88	39.02
		Random	31.24	26.89	31.00
	6	Promoter	3.51	3.34	2.88
		Downstream	4.33	3.19	3.33
		Random	1.57	0.94	1.17
6kb	5	Promoter	78.99	67.46	77.16
		Downstream	80.81	69.28	79.89
		Random	79.25	71.03	74.88
	6	Promoter	39.15	33.10	37.63
		Downstream	43.60	35.12	41.71
		Random	30.93	27.29	30.48

Table 4.2b: Distribution of genes with overlapping coordinates in different chromosomes

Chromosome	Total	% Overlap
Chr 6	3086	22.65%
Chr 7	2891	21.90%
Chr 11	2995	20.50%
Chr 12	2575	24.82%
Chr13	1768	18.27%
Chr 18	1022	22.20%

As another control set, we have also compared the motif vector correlations (Pearson's correlation) in the promoter regions of all genes between chr12-chr17 and chr18-chr19 with Spearman's rank correlation. We observed that the trend was similar for both chromosome pairs 12-17 and 18-19 with R^2 values 0.83 and 0.79 respectively for 1kbps 5mer motif vector comparisons ([figure 4.4](#)).

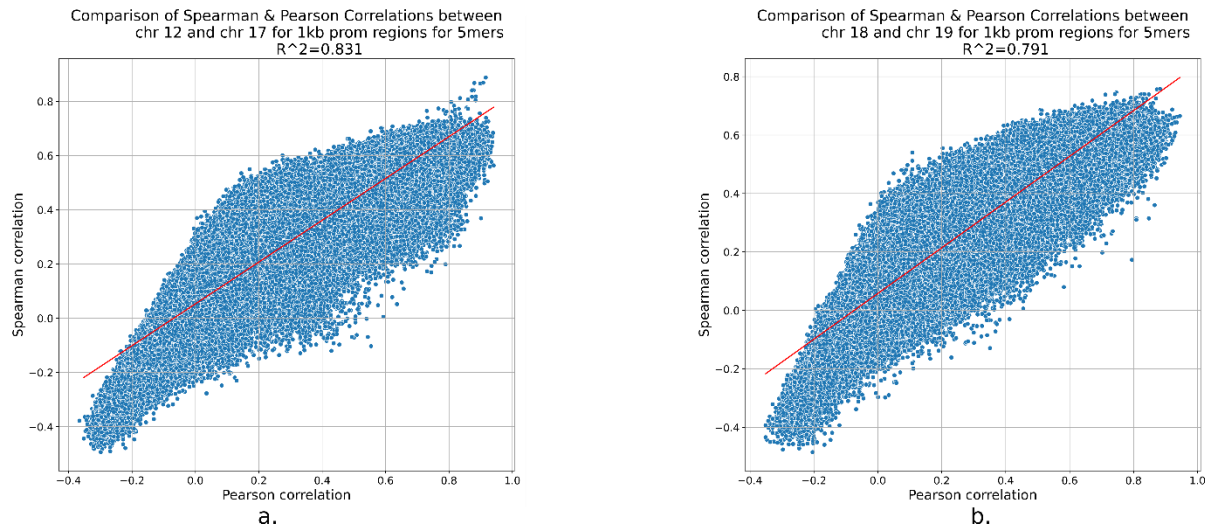


Figure 4.4: Comparison of spearman vs Pearson's correlation coefficient for motif vector correlations of all gene promoters. The x-axis represents Pearson's correlation score and the y-axis represents Spearman's score. Panels a and b show the comparisons for chromosome pairs 12-17 and 18-19 respectively.

4.3.3 Features of gene promoters having high correlations:

We observed from our analysis that high correlations (≥ 0.9) between different chromosome pairs share different features. We compared the sequence similarity between the chromosome pair 13-14 (which is involved in Robertsonian translocation as mentioned in [section 4.3.2](#)) compared to chromosome pair 18-19. We did this comparison with chromosomes 18-19 as the chromosomes are reported to have a staking difference in their gene density⁶⁸ ([figure 4.3](#)) as well as their localization inside the cell nucleus^{69,70}. We compared the sequence identity of the promoter sequence of the highly correlated genes (Pearson's correlation coefficient ≥ 0.9) in chromosomes 13 and 14 against chromosomes 1 and 19. We observed that there is a stark difference in the sequence identity of the promoter sequence of the genes in the two chromosome pairs. Chromosomes 13-17 have 20% of the high correlated genes having a sequence identity of at least 80% ([figure 4.5](#)). Whereas, between the promoter regions of the genes in 18-19 the maximum sequence identity is 37% ([figure 4.5](#)). There are no genes in chromosomes 18-19 that inspire sharing similarity in the motif abundances in the promoter regions, the sequence similarity is not more than 37%.

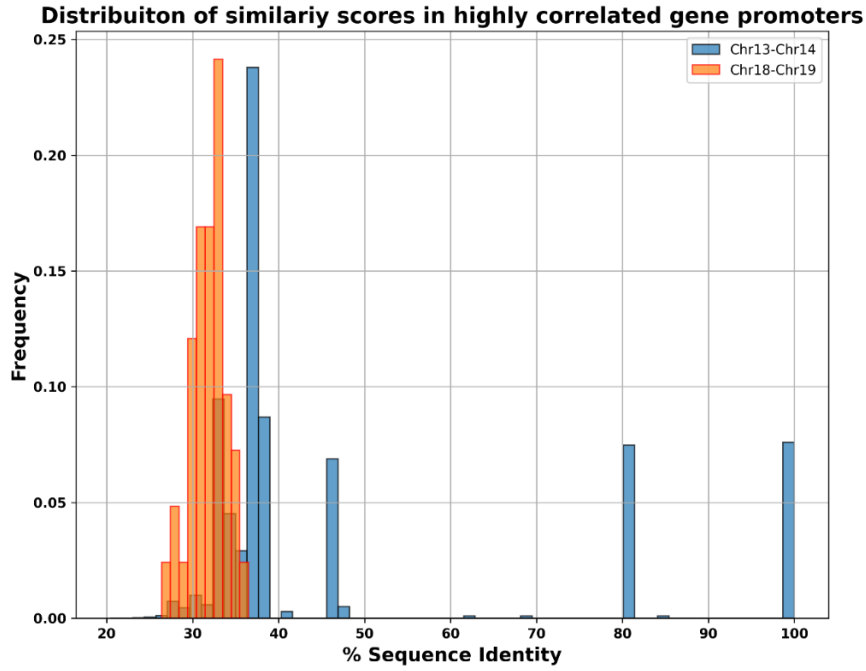


Figure 4.5: Distribution of sequence similarity in gene promoters having high correlation in chromosome pair 13-14 and 18-19 in blue and orange respectively. The x-axis represents the percentage sequence identity in the promoters and the y-axis represents the frequency.

We looked at the relative abundances of the motifs in the gene promoters at different resolutions. This was done to find if the high abundance of certain motifs is a common feature of the gene promoter regions. We wanted to further study, how the motif vector OE_{nkbps_kmer} is distributed for the different promoter and motif sizes ($n = 1, 2$ and 6 kbps and $k = 5$ and 6 -mer). The proportion of genes that have motifs with high OE values is highlighted in [table 4.3](#). Out of the total 63,494 genes annotated across all the chromosomes, we observe that a significant number of motifs show high abundance in the gene promoters across different resolutions. Thus, the high abundance ($OE \geq 5$) holds for a large majority of the genes ($> 50\%$) at different promoters (1kbps, 2kbps and 6kbps) and motif sizes (5 and 6-mers). We also observed that there is also an overlap in these motifs that have high OE in the promoter regions. We identified seven motifs of size 5 (CGGGG, GCGCG, GCGGG, GGCGG, GGGGC, GGGGG and GGAGG) and sixteen motifs of size 6 which have OE_{1kbps_5mer} and $OE_{1kbps_6mer} \geq 10$ in at least 1% of the total genes in each chromosome ([figure 4.6](#)). Thus, our observations highlight certain features of gene promoters. The gene promoters have certain motifs that occur in high abundances which contribute to the correlations when

the motifs are the same. We also highlighted that the high correlation between gene promoters is independent of the sequence identity of the gene promoters.

Table 4.3: Count of gene promoter proximal control elements having motifs with high OE_{nkbps_kmer} . Column 1 represents the resolution of promoter and motif where n is promoter size and k is motif size. Column 2,3,4 shows the result for OE cut-offs 5, 8 and 10 respectively.

Resolution	OE $\geq$ 5	OE $\geq$ 8	OE $\geq$ 10
1kb_5mer	51,103	32,235	22,871
1kb_6mer	57,660	57,534	55,404
2kb_5mer	44,165	21,983	13,370
2kb_6mer	57,659	54,485	48,882
6kb_5mer	33,453	11,412	6,142
6kb_6mer	56,688	45,947	37041

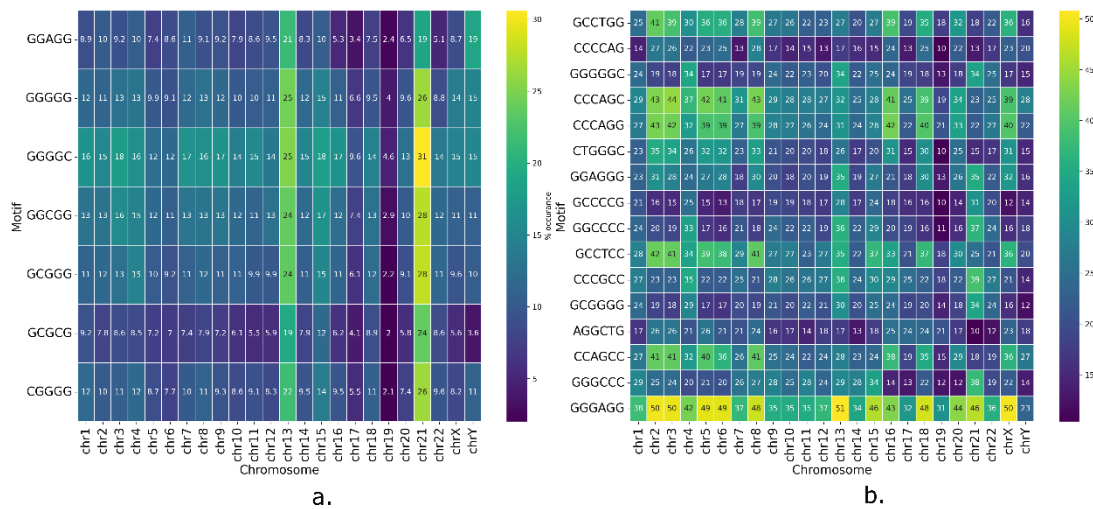


Figure 4.6: Common motifs having high abundance in most gene promoter regions across different chromosomes. Figures 5a and 5b represent the occurrence of motifs of sizes 5 and 6 across different chromosomes. The heatmap scale represents the % of total genes where the motif is present. The lighter shade of yellow represents a larger number of promoters a motif is present. The y-axis represents the conserved motifs and the x-axis represents the different chromosomes. The annotations in the cells represent the percentage of the total genes the motifs are present.

4.3.4 Gene pairs with high correlation share common transcription factors:

To validate whether high correlations within gene pairs located on different chromosomes were attributable to the regulation of these promoters by the same set of transcription factors. On leveraging our promoter scoring approach to identify genes that are associated with a particular transcription factor. By examining the correlation among these genes, we aimed to assess whether the positive correlation scores could be attributed to shared regulation by the identified transcription factor. We looked at 3 genes whose promoter motif vectors had high correlations (≥ 0.9) between one another. The genes identified were LPHN1(ADGRL1) (adhesion G-protein coupled receptor L1) located in chromosome 19, CDK9 (cyclin dependant kinase 9) located in chromosome 9 and TRIM8 (tripartite motif containing 8) located in chromosome 10. All the gene pairs show a strong correlation in the range of 0.9-0.91 among them ([table 4.4](#)). We obtained a high correlation between these motif vectors because there are a few motifs that are common to all three genes with high OE ratio values in the promoter regions ([table 4.5](#)). We used NetworkAnalyst⁵⁸ which allows input of a gene list and provides analysis options, including the identification of common transcription factors using the Jaspar and Encode databases ^{71–74}. Among the many transcription factors associated with these genes, we obtained two transcription factors SP1 and TFAP2A common to all the genes ([figure 4.7](#)). We used JASPAR^{71,72,75} to get the consensus binding sequenced recognized by the transcription factors with our motifs (table 4.5). The consensus for SP1 and TFAP2A can be interpreted as NNGGNNN[G/T][G/C/A][T/A] and GCCNNN[G/A][G/A/T][G/C] respectively. We observed that a subset of the consensus motif (GGCGG, CGGGG, GCCGG) has OE values above 10 in the promoter region of all the genes. Similarly, for TFAP2A, seven 5mer motifs (GCGGC, GGGCG, GCCCG, GGGCC, GGCCG, GCCGG, GGGGC) derived from the consensus have OE values above 10 in all three gene promoters ([table 4.5](#)).

Table 4.4: Correlation of $OE_{prom_1kbp_5mer}$ of LPHN1, CDK9 and TRIM8 genes from chromosomes 19, 9 and 10 respectively.

ChrA	Gene A	ChrB	Gene B	Corr 1kb_5mer
19	LPHN1	9	CDK9	0.905
19	LPHN1	10	TRIM8	0.903
10	TRIM8	9	CDK9	0.910

Table 4.5: Common transcription factors of genes having high correlation and common transcription factors. Column 1 and 2 shows the gene sets and the common transcription factors between them. The consensus from JASPAR for the respective transcription factors is shown in column 3. Column 4 records the $OE_{prom_1kb_5mer}$ of genes.

Genes Connected	TF	Consensus Motif	Promoter Motif: OE																																																																																								
LPHN1,CDK9,TRIM8	SP1		<table> <tr> <th>Motif Pair</th><th>ADGRL1</th><th>CDK9</th><th>TRIM8</th></tr> <tr><td>GGGGC</td><td>17.6</td><td>26.59</td><td>20.83</td></tr> <tr><td>GGCGC</td><td>16.34</td><td>23.95</td><td>23.38</td></tr> <tr><td>CGGCG</td><td>16.33</td><td>29.22</td><td>22.11</td></tr> <tr><td>GGCGG</td><td>16.15</td><td>51.61</td><td>23.49</td></tr> <tr><td>CCGGG</td><td>14.36</td><td>21.28</td><td>25.97</td></tr> <tr><td>GCGGC</td><td>14.35</td><td>41.19</td><td>29.91</td></tr> <tr><td>GGGCG</td><td>14.32</td><td>30.55</td><td>20.83</td></tr> <tr><td>CGGGG</td><td>13.68</td><td>23.91</td><td>27.21</td></tr> <tr><td>GCCGG</td><td>11.74</td><td>15.96</td><td>20.76</td></tr> <tr><td>GGGGG</td><td>11.6</td><td>21.27</td><td>20.54</td></tr> <tr><td>GCGGG</td><td>11.14</td><td>25.19</td><td>20.8</td></tr> <tr><td>GCCCG</td><td>9.15</td><td>23.92</td><td>20.79</td></tr> <tr><td>GGAGG</td><td>8.89</td><td>26.11</td><td>20.29</td></tr> <tr><td>GGGCC</td><td>8.51</td><td>19.93</td><td>14.29</td></tr> <tr><td>GGGAG</td><td>8.31</td><td>14.07</td><td>19.37</td></tr> <tr><td>CGCGG</td><td>7.83</td><td>17.29</td><td>24.68</td></tr> <tr><td>GGAGC</td><td>7.73</td><td>7.52</td><td>9.24</td></tr> <tr><td>AGGCG</td><td>7.71</td><td>10.3</td><td>5.54</td></tr> <tr><td>GGCCG</td><td>7.2</td><td>23.91</td><td>23.37</td></tr> <tr><td>CTCCG</td><td>7.13</td><td>14.97</td><td>7.39</td></tr> <tr><td>GCGCG</td><td>6.53</td><td>10.66</td><td>22.08</td></tr> </table>	Motif Pair	ADGRL1	CDK9	TRIM8	GGGGC	17.6	26.59	20.83	GGCGC	16.34	23.95	23.38	CGGCG	16.33	29.22	22.11	GGCGG	16.15	51.61	23.49	CCGGG	14.36	21.28	25.97	GCGGC	14.35	41.19	29.91	GGGCG	14.32	30.55	20.83	CGGGG	13.68	23.91	27.21	GCCGG	11.74	15.96	20.76	GGGGG	11.6	21.27	20.54	GCGGG	11.14	25.19	20.8	GCCCG	9.15	23.92	20.79	GGAGG	8.89	26.11	20.29	GGGCC	8.51	19.93	14.29	GGGAG	8.31	14.07	19.37	CGCGG	7.83	17.29	24.68	GGAGC	7.73	7.52	9.24	AGGCG	7.71	10.3	5.54	GGCCG	7.2	23.91	23.37	CTCCG	7.13	14.97	7.39	GCGCG	6.53	10.66	22.08
Motif Pair	ADGRL1	CDK9	TRIM8																																																																																								
GGGGC	17.6	26.59	20.83																																																																																								
GGCGC	16.34	23.95	23.38																																																																																								
CGGCG	16.33	29.22	22.11																																																																																								
GGCGG	16.15	51.61	23.49																																																																																								
CCGGG	14.36	21.28	25.97																																																																																								
GCGGC	14.35	41.19	29.91																																																																																								
GGGCG	14.32	30.55	20.83																																																																																								
CGGGG	13.68	23.91	27.21																																																																																								
GCCGG	11.74	15.96	20.76																																																																																								
GGGGG	11.6	21.27	20.54																																																																																								
GCGGG	11.14	25.19	20.8																																																																																								
GCCCG	9.15	23.92	20.79																																																																																								
GGAGG	8.89	26.11	20.29																																																																																								
GGGCC	8.51	19.93	14.29																																																																																								
GGGAG	8.31	14.07	19.37																																																																																								
CGCGG	7.83	17.29	24.68																																																																																								
GGAGC	7.73	7.52	9.24																																																																																								
AGGCG	7.71	10.3	5.54																																																																																								
GGCCG	7.2	23.91	23.37																																																																																								
CTCCG	7.13	14.97	7.39																																																																																								
GCGCG	6.53	10.66	22.08																																																																																								

LPHN1,CDK9,TRIM8	TFAP2A		Motif <table> <tr> <th>Pair</th><th>ADGRL1</th><th>CDK9</th><th>TRIM8</th></tr> <tr><td>GGGGC</td><td>17.6</td><td>26.59</td><td>20.83</td></tr> <tr><td>GGCGC</td><td>16.34</td><td>23.95</td><td>23.38</td></tr> <tr><td>CGGCG</td><td>16.33</td><td>29.22</td><td>22.11</td></tr> <tr><td>GGCGG</td><td>16.15</td><td>51.61</td><td>23.49</td></tr> <tr><td>CCGGG</td><td>14.36</td><td>21.28</td><td>25.97</td></tr> <tr><td>GCGGC</td><td>14.35</td><td>41.19</td><td>29.91</td></tr> <tr><td>GGGCG</td><td>14.32</td><td>30.55</td><td>20.83</td></tr> <tr><td>CGGGG</td><td>13.68</td><td>23.91</td><td>27.21</td></tr> <tr><td>GCCGG</td><td>11.74</td><td>15.96</td><td>20.76</td></tr> <tr><td>GGGGG</td><td>11.6</td><td>21.27</td><td>20.54</td></tr> <tr><td>GCGGG</td><td>11.14</td><td>25.19</td><td>20.8</td></tr> <tr><td>GCCCC</td><td>9.15</td><td>23.92</td><td>20.79</td></tr> <tr><td>GGAGG</td><td>8.89</td><td>26.11</td><td>20.29</td></tr> <tr><td>GGGCC</td><td>8.51</td><td>19.93</td><td>14.29</td></tr> <tr><td>GGGAG</td><td>8.31</td><td>14.07</td><td>19.37</td></tr> <tr><td>CGCGG</td><td>7.83</td><td>17.29</td><td>24.68</td></tr> <tr><td>GGAGC</td><td>7.73</td><td>7.52</td><td>9.24</td></tr> <tr><td>AGGCG</td><td>7.71</td><td>10.3</td><td>5.54</td></tr> <tr><td>GGCCG</td><td>7.2</td><td>23.91</td><td>23.37</td></tr> <tr><td>CTCCG</td><td>7.13</td><td>14.97</td><td>7.39</td></tr> <tr><td>GCGCG</td><td>6.53</td><td>10.66</td><td>22.08</td></tr> </table>	Pair	ADGRL1	CDK9	TRIM8	GGGGC	17.6	26.59	20.83	GGCGC	16.34	23.95	23.38	CGGCG	16.33	29.22	22.11	GGCGG	16.15	51.61	23.49	CCGGG	14.36	21.28	25.97	GCGGC	14.35	41.19	29.91	GGGCG	14.32	30.55	20.83	CGGGG	13.68	23.91	27.21	GCCGG	11.74	15.96	20.76	GGGGG	11.6	21.27	20.54	GCGGG	11.14	25.19	20.8	GCCCC	9.15	23.92	20.79	GGAGG	8.89	26.11	20.29	GGGCC	8.51	19.93	14.29	GGGAG	8.31	14.07	19.37	CGCGG	7.83	17.29	24.68	GGAGC	7.73	7.52	9.24	AGGCG	7.71	10.3	5.54	GGCCG	7.2	23.91	23.37	CTCCG	7.13	14.97	7.39	GCGCG	6.53	10.66	22.08
Pair	ADGRL1	CDK9	TRIM8																																																																																								
GGGGC	17.6	26.59	20.83																																																																																								
GGCGC	16.34	23.95	23.38																																																																																								
CGGCG	16.33	29.22	22.11																																																																																								
GGCGG	16.15	51.61	23.49																																																																																								
CCGGG	14.36	21.28	25.97																																																																																								
GCGGC	14.35	41.19	29.91																																																																																								
GGGCG	14.32	30.55	20.83																																																																																								
CGGGG	13.68	23.91	27.21																																																																																								
GCCGG	11.74	15.96	20.76																																																																																								
GGGGG	11.6	21.27	20.54																																																																																								
GCGGG	11.14	25.19	20.8																																																																																								
GCCCC	9.15	23.92	20.79																																																																																								
GGAGG	8.89	26.11	20.29																																																																																								
GGGCC	8.51	19.93	14.29																																																																																								
GGGAG	8.31	14.07	19.37																																																																																								
CGCGG	7.83	17.29	24.68																																																																																								
GGAGC	7.73	7.52	9.24																																																																																								
AGGCG	7.71	10.3	5.54																																																																																								
GGCCG	7.2	23.91	23.37																																																																																								
CTCCG	7.13	14.97	7.39																																																																																								
GCGCG	6.53	10.66	22.08																																																																																								
LPHN1,TRIM8	E2F1		Motif <table> <tr> <th>Pair</th><th>LPHN1</th><th>TRIM8</th><th>CDK9</th></tr> <tr><td>TTTTC</td><td>3.20</td><td>1.32</td><td>0</td></tr> <tr><td>TTTCC</td><td>2.97</td><td>0.93</td><td>0.47</td></tr> <tr><td>TGCGC</td><td>2.98</td><td>6.46</td><td>0.64</td></tr> </table>	Pair	LPHN1	TRIM8	CDK9	TTTTC	3.20	1.32	0	TTTCC	2.97	0.93	0.47	TGCGC	2.98	6.46	0.64																																																																								
Pair	LPHN1	TRIM8	CDK9																																																																																								
TTTTC	3.20	1.32	0																																																																																								
TTTCC	2.97	0.93	0.47																																																																																								
TGCGC	2.98	6.46	0.64																																																																																								
TRIM8,CDK9	HINFP		Motif <table> <tr> <th>Pair</th><th>CDK9</th><th>TRIM8</th><th>LPHN1</th></tr> <tr><td>ACCCC</td><td>4.7</td><td>3.69</td><td>0.58</td></tr> <tr><td>TCCTC</td><td>6.57</td><td>5.23</td><td>1.63</td></tr> <tr><td>TCCGG</td><td>6.56</td><td>6.46</td><td>2.98</td></tr> <tr><td>AGGGG</td><td>8.44</td><td>8.32</td><td>4.76</td></tr> <tr><td>GTCCG</td><td>1.87</td><td>1.84</td><td>0.6</td></tr> </table>	Pair	CDK9	TRIM8	LPHN1	ACCCC	4.7	3.69	0.58	TCCTC	6.57	5.23	1.63	TCCGG	6.56	6.46	2.98	AGGGG	8.44	8.32	4.76	GTCCG	1.87	1.84	0.6																																																																
Pair	CDK9	TRIM8	LPHN1																																																																																								
ACCCC	4.7	3.69	0.58																																																																																								
TCCTC	6.57	5.23	1.63																																																																																								
TCCGG	6.56	6.46	2.98																																																																																								
AGGGG	8.44	8.32	4.76																																																																																								
GTCCG	1.87	1.84	0.6																																																																																								
LPHN1,CDK9	RELA		Motif <table> <tr> <th>Pair</th><th>LPHN1</th><th>CDK9</th><th>TRIM8</th></tr> <tr><td>TTCCC</td><td>2.16</td><td>1.32</td><td>0.00</td></tr> </table>	Pair	LPHN1	CDK9	TRIM8	TTCCC	2.16	1.32	0.00																																																																																
Pair	LPHN1	CDK9	TRIM8																																																																																								
TTCCC	2.16	1.32	0.00																																																																																								

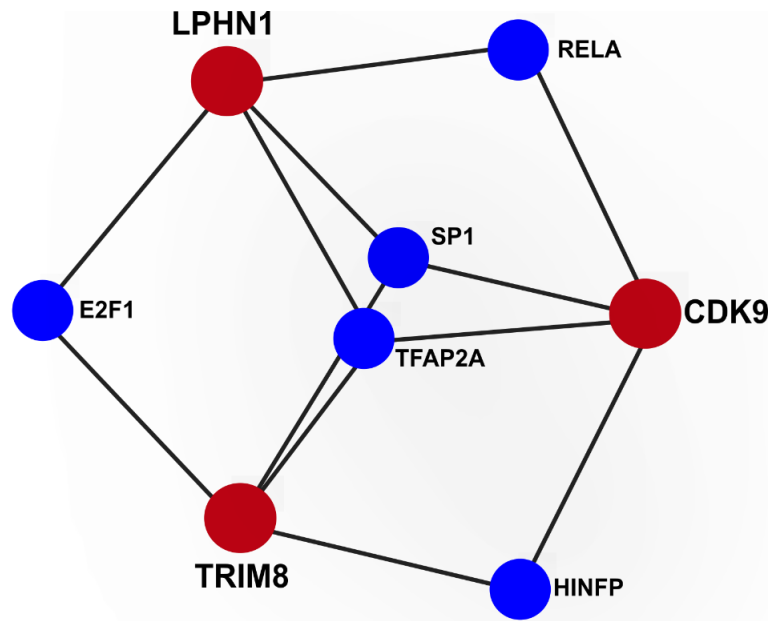


Figure 4.7: A graph representation of a gene regulatory network involving 3 genes LPHN1, CDK9 and TRIM8 (shown as red nodes). Each such red node is connected via edges to transcription factors controlling their expression (cyan nodes).

We also looked at the relative abundances of motifs in the promoters for TFs that were common to two of the three proteins. We looked for motifs that were abundant in the promoters of two of the genes while being poorly represented in the other ([table 4.5](#)). The TFs – E2F1, HINFP and RELA are the TFs that are common between (LPNH1 and TRIM8), (CDK9 and LPNH1) and (CDK9 and TRIM8) respectively. The JASPAR consensus motifs for these TFs can be interpreted as TTT[G/C][G/C]CG[C/G], [A/G]C/G]GTCCGC and [G/C/T][G/T]G[A/G]NTTTCC for E2F1, HINFP and RELA respectively. Though there are TF consensus-derived motifs that have higher OE ratio values in the promoters of the two genes sharing the same TF than the one not in common, the absolute values of the abundances are lower than the other two. We noticed this for all three TFs E2F1, HINFP and RELA where 3, 5 and 1 motifs showed such trend in the relative abundances respectively. This may be indicative of the fact that differential regulation may not always involve TFs alone.

This analysis enabled us to determine if the correlated gene promoters shared a common transcriptional regulator. The identification of common transcription factors supports the hypothesis of shared regulation and suggests the existence of coordinated regulatory mechanisms underlying the observed correlation between gene promoters.

4.3.5 Multiple gene regulation by a common transcription factor:

We previously established that gene pairs sharing motif abundances in their promoter regions are likely to be co-regulated by the same transcription factor(s). We also looked at the promoter regions of all genes that have been reported to be regulated by the same transcription factor. We focussed on the activator protein 1 (AP1) which is composed of dimers consisting of members from the Jun, Fos and activating transcription factor protein families^{76,77}. AP-1 is known to regulate 19 genes by binding to their promoter and enhancer regions, and its function has been extensively studied in mice⁷⁸. These genes have been reported to be upregulated or downregulated by the AP-1 family of proteins in mice⁷⁸. We looked at the human homologous of the mouse genes and looked at the motif distribution in their promoter regions. All vs all correlations of these 19 gene promoter motif vectors were computed ([figure 4.8](#)). We observed that some of the genes share higher correlations with other gene promoters. Clustering based on the degree of correlations, we observed three groups based on the genes which have higher correlations within themselves compared to others in the list of genes. Our analysis revealed that 10 out of the 19 selected genes exhibited a strong positive correlation (≥ 0.5) with their human counterparts. An interesting observation here is that the Fos/Jun binding consensus motif TGACTCA is not among the motifs with high OE ratio values. There are motifs with high OE (OE ratio ranging from 3 to 33) that are in common with several genes (such as motif CGCGG present in genes BCL2, BCL2L11, HBEGF and EGFR) and could be binding sites of some gene regulators (identity unknown) (supplementary data of AP1 network). This finding suggests a conservation of regulatory mechanisms in the promoter regions of these genes between mice and humans. The observed positive correlation scores further support the notion that these genes are likely part of the same regulatory machinery controlled by the AP-1 transcription factor. This also highlights how we can generate a gene regulatory network by clustering genes whose promoters share the same consensus motif.

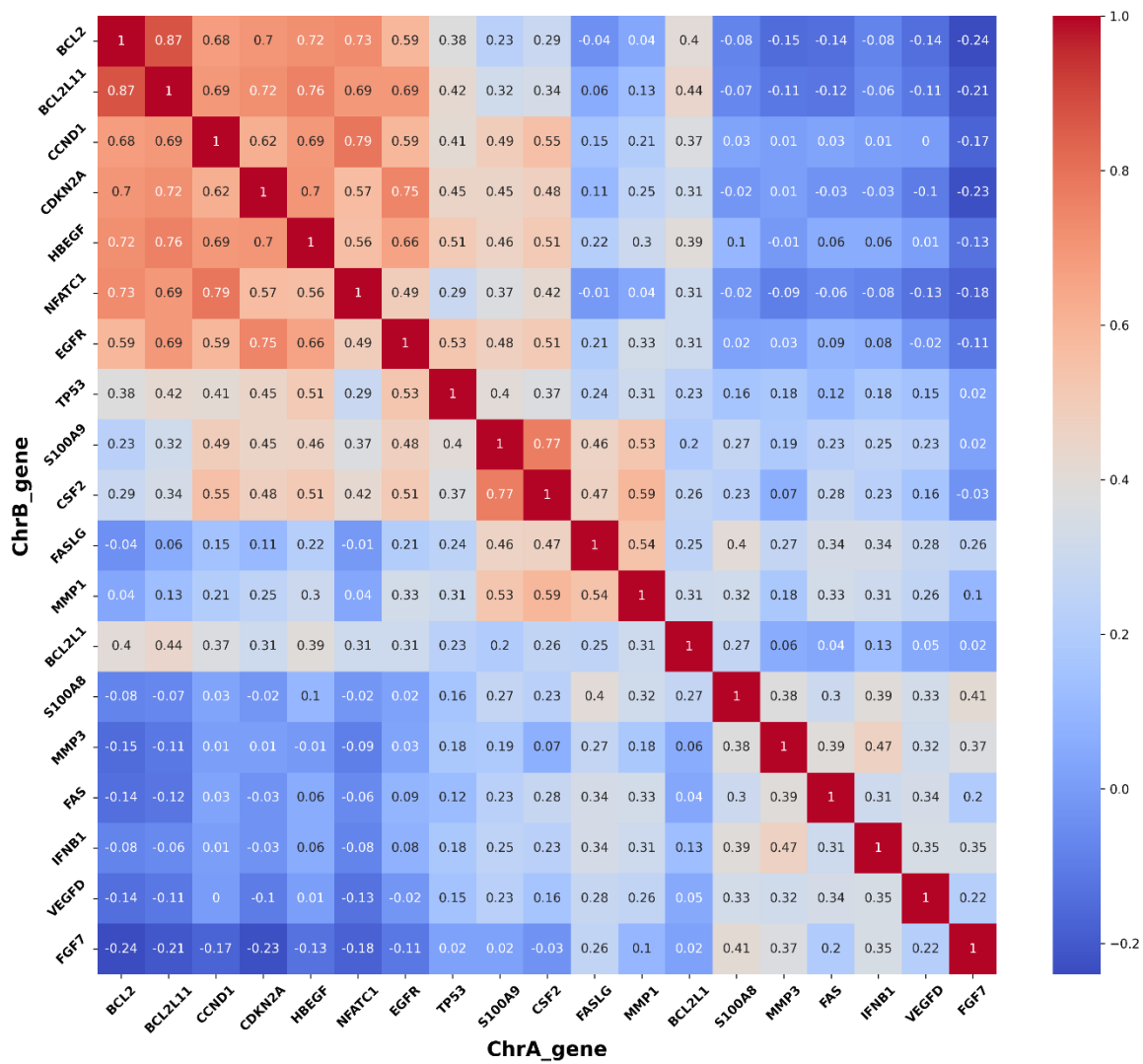


Figure 4.8: A gene-gene interaction heat map of OE_{1kb_5mer} for 19 genes that are all regulated by the transcription factor Jun-Fos. The higher (or lower) the correlation the darker the shade of red (or bluer) is the grid.

4.3.6 Promoter correlations highlight spatial proximity:

We previously established that genes whose promoters share high correlations are likely to be co-regulated. Thus, we investigated the spatial proximity of such genes. We wanted to study whether the correlations implied co-localization. Experimentally, co-localization information is inferred using Hi-C^{18,26} experiments which used chemical cross-linking to identify genomic regions which are in proximity to one another. The resolution from different Hi-C experiments is in the range of 10-100 kbps. In contrast, our computations are at a higher resolution, with our inferences being drawn by studying the abundance in the size of 5-6 bps motifs. Hi-C experiments capture the

cross-links, independent of the presence or absence of genes. Our data is restricted to gene promoters and thus focuses only on the genic regions. Thus our analysis is in comparison with only a subset focusing only on the Hi-C contacts which have a gene in them. Additionally, we also looked at the promoter capture Hi-C data (pcHi-C) which captures distal promoter interacting regions for all promoters⁵. We compared our correlations of the gene pairs that overlap with the spatially proximal; regions detected by the Hi-C experiments.

We used two datasets for our validation: 1. Hi-C data of long-range chromatin contacts in HCT116 colon cancer cells ⁶⁰(accessible under GEO accession number GSE18215) and 2. pcHi-C data which looks at novel gene contacts associated with autoimmune risk loci ⁶¹ (accessible under GEO accession number: GSM1704495). We compared our motif vector correlations to Hi-C data for chromosome pairs 18 and 19. Chromosomes 18 and 19 have 1022 (spread over ~80.5mbps) and 2531 genes (spread over ~61.7mbps) respectively according to the current T2T reference genome assembly (refer to table 6 for annotations of the previous reference assemblies). The two chromosomes differ in their gene density⁶⁸ ([figure 4.3](#)) as well as their localization inside the cell nucleus^{69,70}. Chromosome 18 has a peripheral localization whereas chromosome 19 is mostly internalized. We obtained the genome coordinates of contacts from the Hi-C experiments and obtained the genes that overlap with the Hi-C contacts. Identification of the gene coordinates was done using the respective reference assembly ([table 4.6](#)) that was used to generate the Hi-C data.

From the 10kb bin long-range chromatin contact Hi-C data, there are Hi-C data, there are 1,327,617, 699,290 and 23,688 Hi-C contacts intra-chromosome-18, intra-chromosome-19 and inter-18-19 respectively according to GRCh38 reference assembly. Of these, only a small fraction of 12,790, 115,595 and 926 are contacts overlap with regions having genes in intra-18, intra-19 and inter-18-19 respectively. As expected the number of inter 18-19 contacts is small in comparison to the intra chromosome (18-18 and 19-19) contacts. The aim here was to determine if the gene promoter motif vector correlations correlate with physical proximity. Of the overlaps determined, only 12,213, 111,730 and 890 gene-pairs find a match with the current T2T gene annotation list which is our reference for the motif vector correlations.

Table 4.6: Comparison of number of genes annotated according to different reference assemblies

Chromosome	T2T	GRCh38	GRCh37
Chr18	1022	983	717
Chr19	2531	2492	2248

Table 4.7: Comparison of gene-pair correlation inter and intra chromosome 18 and chromosome 19 in Hi-C contacts vs all

Chr-Pair	Resolution	1kb_5mer	1kb_6mer	2kb_5mer	2kb_6mer	6kb_5mer	6kb_6mer
Chr18-Chr19	Gene pairs with motif vector correlation ≥ 0.5	9.92%	0.47%	32.09%	2.32%	76.20%	38.90%
	Gene pairs in HiC with motif vector correlation ≥ 0.5	14.38%	1.01%	35.96%	4.16%	77.75%	44.38%
Chr18-Chr18	Gene pairs with motif vector correlation ≥ 0.5	9.96%	1.18%	31.94%	2.31%	80.11%	37.28%
	Gene pairs in HiC with motif vector correlation ≥ 0.5	18.84%	7.72%	41.8	8.95%	88.85%	52.25%
Chr19-Chr19	Gene pairs with motif vector correlation ≥ 0.5	15.10%	0.90%	46.07%	9.14%	79.56%	60.74%
	Gene pairs in HiC with motif vector correlation ≥ 0.5	22.18%	3.74%	56.91%	15.11%	88.30%	72.78%

We identify the genes having motif vector correlation scores ≥ 0.5 from the inter and intra chr18, chr19 Hi-C contacts. We see enrichment in the motif vector correlations in the Hi-C contacts for all the inter and intra-Hi-C contacts at all resolutions of motif vector correlations at different sizes ([Table 4.7](#)). However, we see that the enrichment improves if we are looking at promoters in very close proximity to the gene transcription start sites (1kbps). We also verified the high correlations between the genes as a result of sequential proximity between the genes. As a control, we plotted the distance between the genes against the correlation of their promoter control regions. We observe that high correlations between gene promoter proximal control regions are not dependent on the sequential proximity of the genes ([figure 4.9](#)). It is to be noted that this comparison was done only across intra-gene correlations as we cannot determine the sequential proximity of inter-chromosome correlations.

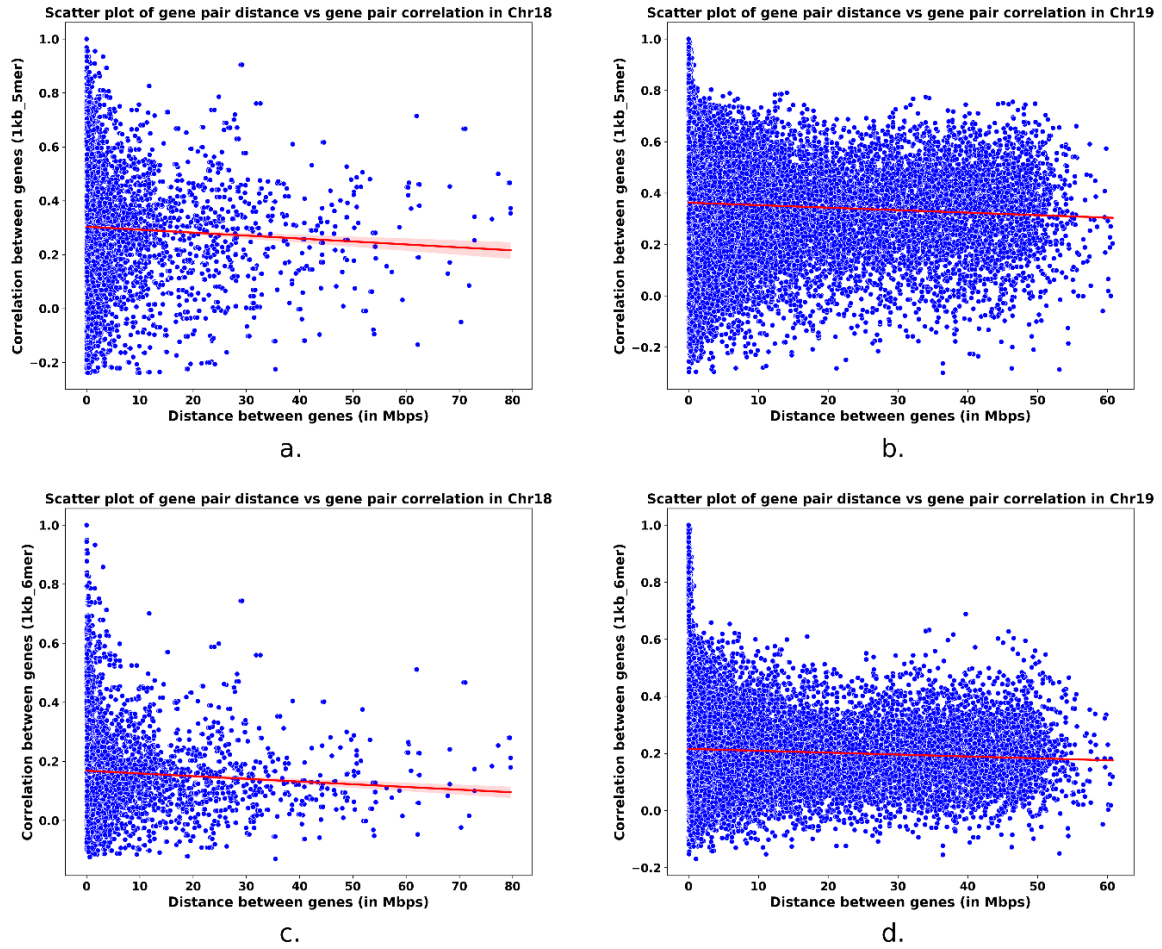


Figure 4.9: Comparison of distance between genes and motif vector correlations.

The y-axis represents the correlation coefficient and the x-axis represents the distance between a gene pair. Panel 7a and 7b show motif vector correlations at 1kb 5mer for gene pair in chromosomes 18 and 19 respectively. Panels 7c and 7d show the correlations at 1kb 6mer resolutions in the same order.

We repeated this analysis using pHi-C data using chromosome pairs 18 and 19 and obtained 3655 contacts. Out of this when we obtained the coordinates of gene pairs using the GRCh37 reference assembly annotations, we found that only 123 gene coordinates overlap with the pHi-C contacts. 118 out of the 123 gene-pairs match with our correlations using current T2T reference assembly gene annotations ([table 4.8](#)).

Table 4.8: Comparison of gene-pair correlation inter chromosome 18 and chromosome 19 in Hi-C vs pHi-C vs all

Resolution	1kb_5mer	1kb_6mer
Gene pairs with motif vector correlation ≥ 0.5	9.92%	0.47%
Gene pairs in HiC with motif vector correlation ≥ 0.5	14.38%	1.01%
Gene pairs in pHiC with motif vector correlation ≥ 0.5	22.80%	0.85%

TADs are well-defined chromosome segments which make more frequent genomic contacts to one another and are hence purportedly in spatial proximity⁷⁹. It has been established that binding sites for the protein CTCF, characterised by the motif CCCTC, mark the TAD boundaries^{80,81}. The CCCTC motif is among the top 50 most abundantly present motifs in the human genome (43rd on a list of 512 with an average OE_{1kbps_5mer} value of 1.85. Interestingly, the OE_{1kbps_5mer} for CCCTC in the promoter regions of gene pairs that have a motif vector correlation of > 0.5 and are proximal to Hi-C or the pHi-C contacts in chromosomes 18 and 19 lies in the range 0 to 20 ([figure 4.10](#)). In contrast, the OE_{1kbps_5mer} for the other motifs in the genes in proximity, do not exceed 2.5. On closer inspection, we observed that 10 and 16 out of the 27 and 128 gene pairs in pHi-C and Hi-C respectively show an OE_{1kbps_5mer} of the CTCF motif greater than 5 in at least one of the genes in contact. Thus, the CTCF motif is one of the major contributors to the high correlation (correlation coefficient > 0.5) in the genes in proximity.

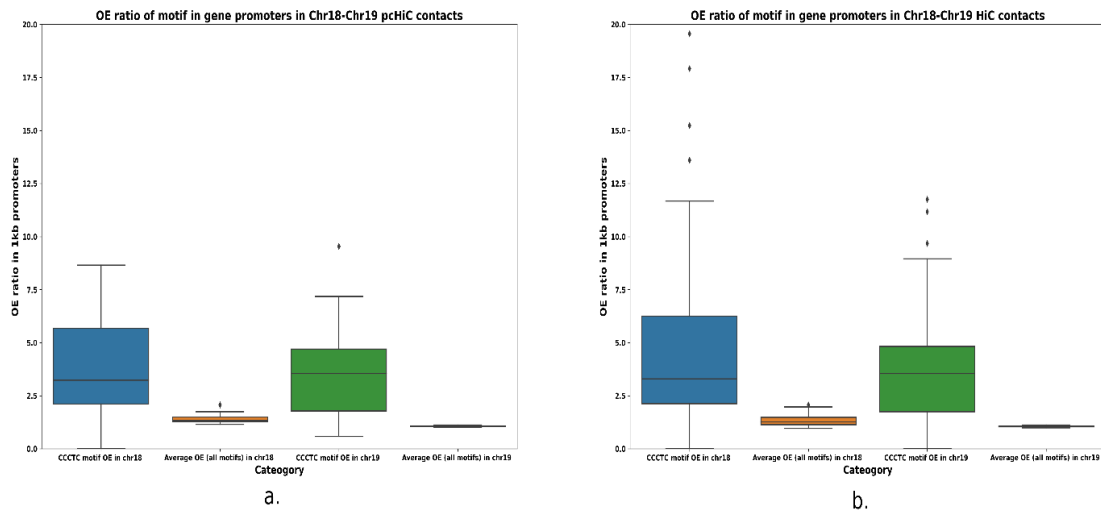


Figure 4.10: Distribution of OE_{1kbp_5mer} CTCF binding motifs vs other motifs in gene promoters with Hi-C and pcHi-C contacts between chromosome 18-19 Hi-C. Panels a and b represent the distribution of genes from pcHi-C and Hi-C experiments respectively. The blue and green colour box plot represents the distribution of the CCCTC motif chromosome 18 and chromosome 19 respectively. Orange and black colour represent the distribution of other motifs in chromosome 18 and chromosome 19 respectively.

We see the enrichment conserved when compared to all gene correlations. The motif vector correlations are high among gene pairs that also have Hi-C contacts. This opens up the possibility that Hi-C contacts could be predicted from gene correlations. It is to be noted here that the gene pairs in 10kbps bin Hi-C contacts do not overlap with the pcHi-C data even though the individual genes from chr18 and chr19 show overlap ([figure 4.11](#)). Due to the lack of uniformity in the reference assembly used in generating the Hi-C data, we believe that we are missing out on identifying many more genes in the Hi-C data that are captured by our correlations because the GRCh37/ GRCh38 annotations and gene positions are not uniform with the T2T annotations.

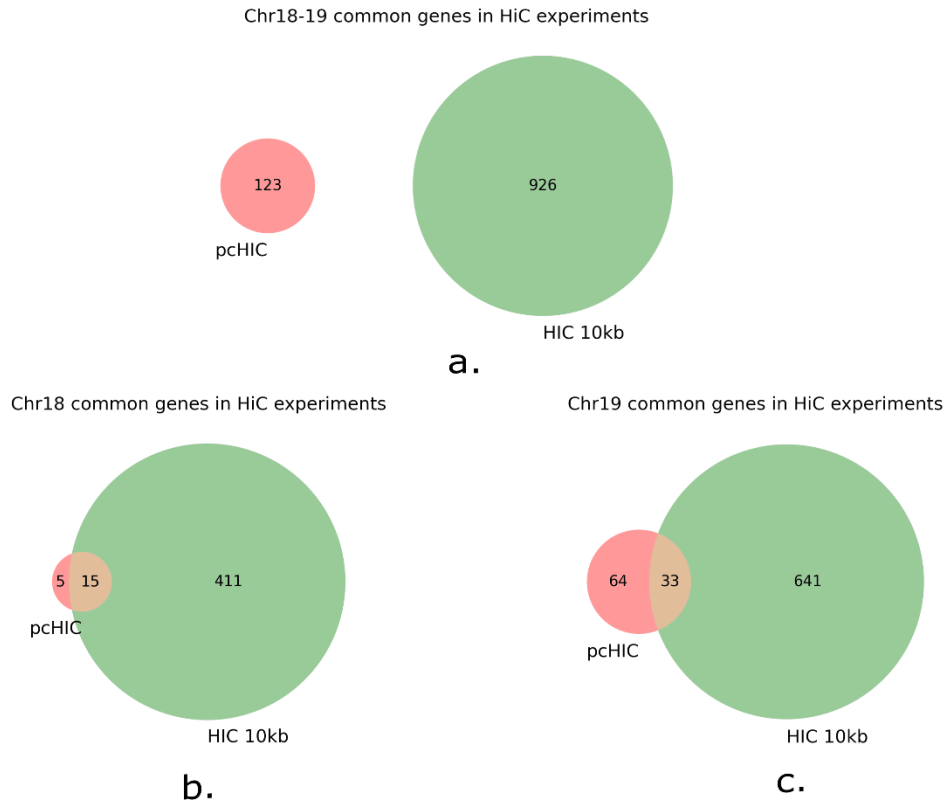


Figure 4.11: Overlap of gene/gene-pairs in Hi-C vs pHi-C data. 5a. represents the overlap between all the gene-pair. Figures 5b and 5c represent the overlap between the unique genes involved in different Hi-C contacts in Chr18 and Chr19 respectively. The green colour represents Hi-C capture at 10kbps resolution and the peach colour represents pHi-C contacts.

4.4 Discussion:

The findings from our comprehensive analysis of motif distributions and promoter correlations shed light on the regulatory landscape of the genome and reveal potential coordination among different chromosomes. The observed patterns of motif representation across the genome, as depicted by the heatmaps, indicate the existence of common motifs that are consistently overrepresented, independent of the ATGC content. These results suggest the presence of underlying regulatory mechanisms that influence the distribution and conservation of specific DNA motifs. The increasing refinement of motif patterns as motif size increases highlights the potential functional significance of larger motifs in regulatory processes. The identification of overrepresented motifs ranging from dinucleotides to 6-mers suggests that longer motifs may play a more prominent role in governing

gene regulation. This underscores the importance of considering larger motifs when studying the functional implications of DNA sequence patterns.

The examination of promoter regions across chromosomes provided valuable insights into shared regulatory mechanisms and potential interactions among genes. We observed that the specific 5 and 6-meric motifs had higher abundance in the promoter regions (~50%). When the correlation was obtained based on the motif abundances, we obtained certain gene promoters had strong correlation (≥ 0.9) compared to others. We obtained two such clusters of high correlations between genes in different chromosomes: (13, 14, 15, 21 and 22) and (4,10). In both these groups the chromosomes are reported to be involved in translocations. Chromosomes 13,14,15, 21 and 22 show a mix and match of chromosomal segments between their genes is termed Robertsonian translocations and is implicated in genetic diseases such as Patau syndrome⁸² and Down syndrome⁸³. Translocations between chromosomes 4 and 10 are associated with certain types of leukaemia^{65,66} and Wolf-Hirschhorn syndrome⁶⁷. The motif vector correlations also highlight the trends. The high abundances of promoter correlations are also an attribute of the tandem repeats in the promoter regions of the rRNA genes across these chromosomes, however this observation is restricted to the chromosome group of (13,14,15,21 and 22). We also highlighted that the translocations may only be possible when there is a high degree of sequence similarity/identity between the regions of translocation. However, not all regions sharing high similarity/identity may be susceptible to translocations as it may require the presence of certain motif(s). What motif(s) and their distributions can only be gauged with a focused analysis of the translocating segments. The patterns of high correlation across gene promoters are also distinct from that of downstream regions of genes or random regions in the chromosomes.

To validate the reliability of our promoter scoring metric, we conducted three distinct case studies. In the first study, we randomly selected a list of genes with high correlation scores and performed a NetworkAnalyst analysis. This analysis revealed common transcription factors between SCAF1 and other genes, providing evidence of shared regulation. Additionally, we identified a larger regulatory network involving multiple genes based on the common transcription factors shared between different gene pairs.

The second validation study focused on the regulatory role of the AP-1 transcription factor. By examining the promoter regions of genes regulated by AP-1 in mice, we found a strong positive correlation in their human counterparts, indicating the conservation of regulatory mechanisms. These results support the involvement of AP-1 in coordinating gene expression across species and highlight the importance of regulatory networks involving these genes.

In the third study, we compared our results with Hi-C and pcHi-C data. We selected chromosome pairs 18 and 19 which have opposite spatial positing in the cell nucleus. Chromosome 18 is found towards the periphery whereas chromosome 19 is found towards the centre. We observed that there is a significant enrichment of high gene correlations in both the Hi-C and the pcHi-C contacts bins that have genes. The genes also have a high abundance of the CTCF binding motifs which are reported to be a marker of the TAD boundaries. This validates the spatial conservation of functionally relevant gene promoters and the potential regulation by common factors. These highlight the potential of our approach to uncover functional connections within the genome. The dataset generated from our analysis serves as a valuable resource for validating known gene networks and generating predictions for novel gene regulatory interactions.

This study contributes to our understanding of gene regulation and the coordination of gene expression across chromosomes. The observed patterns of motif distributions and promoter correlations provide valuable insights into the regulatory landscape of the genome and the potential coordination among genes. This study sets the stage for further investigations into the intricate mechanisms governing gene expression and coordination in the genome.

Overall, our comprehensive analysis of motif distributions and promoter correlations has revealed important insights into the regulatory landscape of the genome and the coordination of gene expression across chromosomes. The presence of consistently overrepresented motifs, independent of the ATGC content, suggests the existence of underlying regulatory mechanisms influencing the distribution and conservation of specific DNA motifs. The increasing refinement of motif patterns with larger motif sizes underscores the potential functional significance of these motifs in gene regulation.

The examination of promoter regions has provided valuable insights into shared regulatory mechanisms and potential interactions among genes across different

chromosomes. The observed higher correlations between certain chromosome pairs suggest the involvement of common transcription factors that influence the regulation of multiple genes. Notably, the absence of overlapping gene promoters in specific chromosomes indicates the influence of structural rearrangements, such as Robertsonian translocations, on gene promoter interactions and correlation patterns.

The validation of our promoter scoring metric through three distinct case studies has further confirmed its reliability and usefulness. The comparison with Hi-C data revealed spatial conservation of functionally relevant gene promoters and the potential regulation by common factors. NetworkAnalyst analysis identified common transcription factors between genes with high correlation scores, indicating shared regulation and the presence of larger regulatory networks. The conservation of regulatory mechanisms in the promoter regions of genes regulated by the AP-1 transcription factor reinforces its importance in coordinating gene expression across species.

In conclusion, our study advances our understanding of gene regulation and highlights the coordinated expression and regulation of genes across chromosomes. The observed motif distributions and promoter correlations provide valuable information for future studies exploring gene regulation and predicting gene regulatory networks. The generated dataset serves as a valuable resource for validating known gene networks and uncovering novel gene regulatory interactions. Further investigations into the intricate mechanisms governing gene expression and coordination in the genome will enhance our understanding of the complex regulatory processes underlying cellular functions.

Chapter 5

Motif-associated gene networks

- [Finding gene associations across chromosomes](#)
- [Establishing gene associations for a single gene of interest](#)
- [Generating a network of multiple genes based on motif abundance](#)

5.1 Introduction:

A gene network is representative of a collection of genes and the associations between them based on different parameters⁸⁴. These networks help us identify the regulatory mechanisms and the molecular interactions that control cellular functioning. The currently updated reference assembly (T2TCHM13) of the human genome has annotated 63,494 genes^{10,85}. It is essential to have tools that classify the genes and find connections between genes based on how they are regulated by the protein binding partners. Looking at the sequence information in the genes or the gene promoter alone does not give information about the binding sites for proteins. This is because specific recognition by proteins occurs at a resolution of 5-6 bps. Using the sequence information from the promoter regions of the genes, we computed the observed to expected (OE) ratio of motifs across all gene promoters. We used the motif vector of the OE ratios to generate a correlation score that establishes the degree of association between a gene pair. We generated a database of these correlation scores. Our server Motif Associated Gene NETworks in Chromosomes, used the correlation score from the database to generate gene networks across different chromosomes ([figure 5.1](#)).

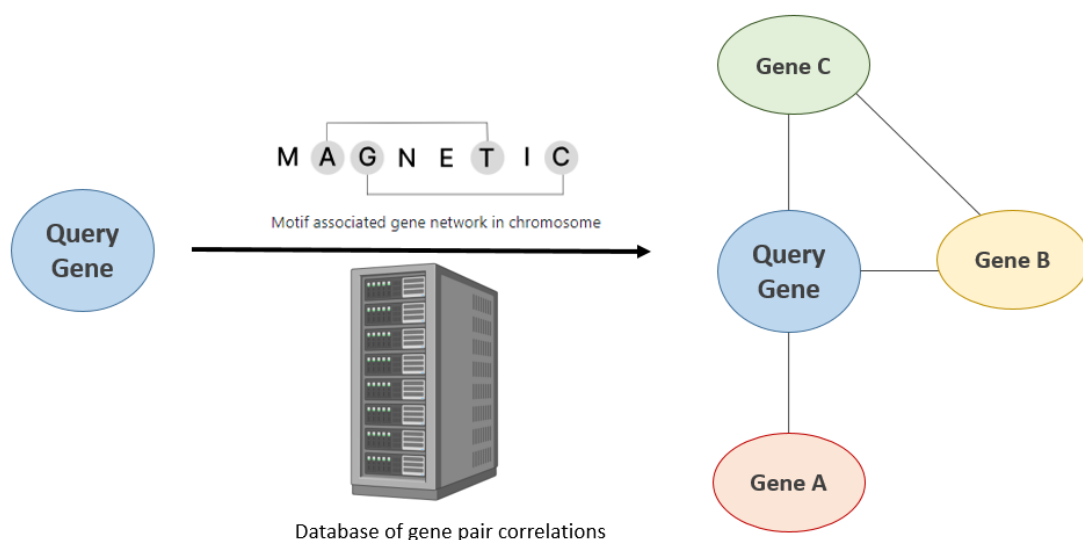


Figure 5.1: Schematic showing the working of the MAGNETIC webserver

The MAGNETIC algorithm looks at the promoter regions of the different annotated genes of 1,2 and 6kbp upstream of the transcription start site (or gene promoters)

and looks at the abundance of different 5 and 6-mer motifs in these regions. Based on the similarity of motif abundances in the promoter regions, a correlation score is computed to associate them. The correlations are obtained for motifs of sizes 5 and 6 in the promoter regions 1,2 and 6kbps upstream of gene start site. In general, any gene pair across different chromosomes are connected at a different correlation score. A high positive correlation score (≥ 0.5) between a gene pair indicates the confidence of the connection based on the repeated occurrence of the same motif (or group of motifs).

The webserver can be used for four purposes: 1) For a query DNA motif, find the genes promoters where the motif is abundance is high 2) For a given gene of interest, find all the genes associated with it across different chromosomes (including the genes in the same chromosomes) 3) For a given query gene list, establish a network of gene connections across chromosomes 4) For a given chromosome pair find the genes that share similar motif distribution. All of the above data is generated based on the correlation of the motif vector in the promoter regions across different lengths.

5.2 Methodology:

The MAGNETIC server generates a network of genes using the abundance of motifs in the promoter regions of the genes. Using this information of motif abundance, we have a pre-computed database of all annotated gene correlations, which is used to generate the network. The calculations for correlations are done based on two parameters: 1) promoter size and 2) motif size. The correlations between all possible gene pairs annotated for a given organism are pre-computed in our database for promoter sizes 1kbps, 2kbps and 6kbps of motif vector resolutions of 5- and 6-mer. The heatmaps, graphs and networks in the webserver are designed using the D2.js module of the JavaScript library. The web framework is designed using Django⁸⁶.

5.3 Program overview:

5.3.1 Input:

The webserver is freely accessible at <http://cospi.iiserpune.ac.in/magnetic> without login requirements. The first step will be the selection of the organism of interest by the user.

Following the selection of the genome of interest, the user will be directed to the options of generating single gene or multiple gene networks. The input is a single or multiple gene name(s) in the HUGO gene nomenclature⁸⁷ (<https://www.genenames.org/tools/search>). In addition to the gene name, the users need to specify the resolution and the correlation cutoff at which the gene network will be generated. The gene resolutions for calculation of the promoter correlations are 1,2 and 6kbp when looking at the motif vectors of size 5- and 6mers. The correlation cutoff values range from 0.0 to 1.0.

5.3.2 Output:

The web server provides three options for generating the associations between genes: 1) Getting all gene associations for a given chromosome pair based on a cutoff of the correlation coefficient. 2) Given a gene of interest find all the genes across different chromosomes associated with it. 3) Given a multiple list of genes, generate a gene network across all genes in different chromosomes. The users can also get information on the occurrence of the most abundant motifs in terms of OE ratio in the promoter regions of the different genes (dependant on motif and promoter sizes). The user will also get information about the common motifs with a highest OE ratios for a given gene pair.

5.4 Server description:

5.4.1 Finding motif abundances across genes:

The MAGNETIC server allows the user to find the motif abundances across different genes. The user gives a query motif of size 5 or 6 and on selecting the promoter resolution the webserver generates a table of the count of genes across different chromosomes that have the motif in the different OE ranges. On mentioning the lower and the upper ranges of the OE scores, the list of genes information along with the OE scores is displayed on a table. The table displays only 20 genes but the remaining genes can be downloaded as a .csv file. For example, in [figure 5.2](#) for query gene TATAT the genes 11 genes that have OE values in the range of 35 to 42. This option will help the user to identify genes promoters harbouring a query motif.

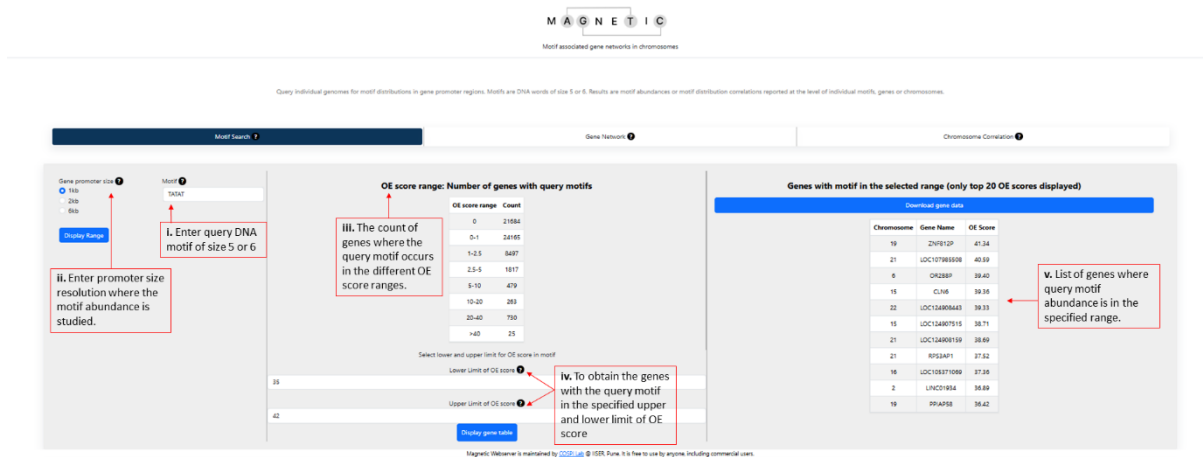




Figure 5.3: A snapshot of the server showing gene associations for a given query gene. Panel a shows the submission of the query using the option in the server. Panel b shows the number of genes per chromosome associated with the query gene. Panel c shows the number of genes per chromosome associated with the query gene. The gene connections above the given cutoff are shown. The different colours for a gene represent the chromosome the gene is located. Panel d. Shows a table of motifs with the highest OE for a highlighted gene association (WNT4). The motifs are followed by the OE ratio in brackets. The motifs highlighted in red are the common high-scoring motifs for the selected gene pair.*

5.4.3 Generating network of multiple genes based on motif abundance:

The MAGNETIC web server can also be used to find correlations between a query gene list using the third option of “Multi gene promoter networks”. Based on the resolution of motifs and promoter sizes, the correlation values are extracted from our database. An example case of generating a gene network using three query genes ADGRL1 (Adhesin g-protein coupled receptor 1), CDK9 (Cyclin dependant kinase 9) and IL1A (Interleukin 1 alpha) is shown in [figure 5.4](#). ADGRL1 is characterised as adhesion G-protein coupled receptor that is reported to be involved in synaptic development, maturation and activity⁹¹. Interleukin 1 alpha belongs to interleukin 1 family and is involved in the inflammatory and the regulation of immune response⁹². On submission of the query genes, the user gets the correlation information of query genes (as shown in panel ii of [figure 5.4](#)). Following the matrix, the user can opt to see the individual connections for all the genes at the selected cutoff to find the overlap of gene connections among the query gene list ([figure 5.4](#) panel iv).



Figure 5.4: A snapshot of the server showing a multiple-gene network. The different panels labelled in red boxes highlight their function. The gene names in bold represent the query genes.*

Using the multiple gene network option of MAGNETIC webserver, we generated a gene network using the query genes. The user can obtain the motif with highest abundance across all the three gene promoters at different resolutions of promoter size. The user can also retrieve the common genes that share high correlation across all or a single query gene. At the given cutoff criteria of 0.9 we observed that all the genes share an association based on the motif abundance. From network analyst⁹³ we cross verified the presence of a shared transcription factor SP1 between the three genes. The role of SP1 and the common transcription factors between ADGRL1 and CDK9 have been discussed in details in [section 4.3.4](#). ILA1A shared common transcriptions factor E2F1 with ADGRL1 and NFKB1 with LPHN1. The MAGNETIC server also gives the user the other gene associations that overlap with the query genes at the given cutoff.

5.4.1 Finding gene associations across chromosomes:

The MAGNETIC server allows the user to look for all gene associations across all chromosomes. The gene associations are determined by a correlation score for a given gene pair. The number of connections between gene pairs is determined based on the cutoff score given by the user. For example, for a cutoff of 0.9, and a resolution of promoter size 1kb and motif size 5mer we obtain maximum connections between chromosomes 13,14,15, 21 and 22. On clicking the cell corresponding to the

chromosome pair, the user will be directed to download the detailed correlations of all the gene pairs between the two chromosomes at the selected cutoff ([figure 5.5](#)).

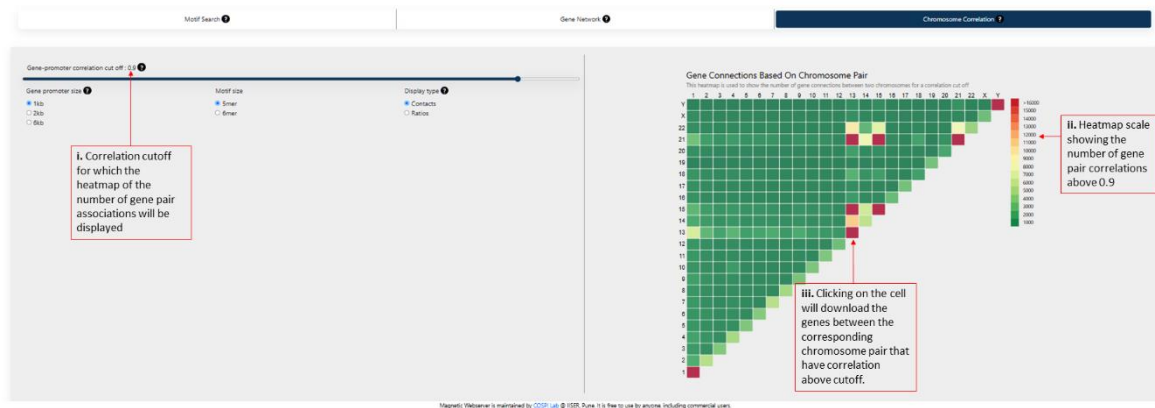


Figure 5.5: A snapshot of the server showing the heatmap of the number of gene-pair correlations above a cutoff of 0.9 across all the chromosome pairs. The colour scale varies from green to yellow to red. The darker the shade of red, the more the number of connections between a given chromosome pair. The red boxes with labels highlight the significance of each section of the server.*

Acknowledgement

Webserver design, maintenance and execution was done by Rohan Topno, Ajinkya Ranshur and Saksham Shrimali

*The snapshots from the webserver referred to in this chapter is subject to change as the webserver is currently being optimised. Broadly the visual description and the information retrieved from the webserver will be the same.

Chapter 6

Role of motif clustering in chromosome positioning

- [Distribution of LAD-iLAD regions across different chromosomes](#)
- [Identification of motifs enriched in LAD and iLAD regions](#)
- [Motifs enriched in LADs are AT rich](#)
- [Identification of evolutionarily conserved motifs enriched in LAD and iLAD regions](#)
- [Prediction of LAD-iLAD regions using Hidden Markov Model](#)
- [Identifying protein targets associated with LADs](#)

6.1 Introduction

The human cell nucleus comprises 46 tightly packed chromosomes of various genomic sizes and gene densities folded in different scales. The packing of the genome in the 3D space is a non-random event. The early work of Carl Rabi and Theodor Boveri proposed the concept of chromosome territory (CT). This demonstrated how chromosomes follow a definite spatial positioning radially (i.e., centre-to-periphery) enabling sequestering specific chromosomal domains to the nuclear envelope⁹⁴. Later on a landmark in the field of genome architecture was the work of Thomas Cremer which provided compelling evidence backing this concept⁹⁵. This mapping of the CTs inside the cell nucleus is aided by numerous factors that involve the participation of sub nuclear compartments involving chromatin-chromatin interactions and other nuclear proteins. One such key player in the closest vicinity of the interphase chromosome is the nuclear lamina (NL) lining the inner nuclear membrane ([Figure 6.1a](#)). The nuclear lamins constitute the major component of the NL⁹⁶.

Lamins are type V intermediate filaments (IFs) that form a mesh-like structure beneath the nuclear envelope. Like other IFs, they have a short globular N-terminal (head) domain, a coiled α -helical rod domain and a carboxyl-terminal tail domain ([figure 6.1b](#))⁹⁷. Based on the biochemical nature and structure, lamins have two major isoforms: Lamin A and B. While the A-type lamins have a neutral isoelectric point, the B-type lamins are more acidic⁹⁸. The amino terminal head and the central rod domain of both A and B-type lamins resemble each other but the A-type lamins have a unique 90 amino acid segment in their expanded carboxyl-terminal domain. The localisation of the lamins is also different. The transient disassembly of the nuclear lamina during mitosis leads to the solubilisation and the dispersal of the A-type lamins whereas the B-type lamins remain associated with the membrane⁹⁹. Lamins A and C are derived from a single gene LMNA by alternate splicing and they also show sequence similarity up to the first 566 amino acids. Lamin C has a shorter sequence and lacks a stretch

of 98 residues that are present in lamin A before the post-translational processing and instead contains a carboxyl terminus of 6 unique amino acids. The other isoform Lamin B is encoded by LMNB1 and LMNB2. The CAAX box, which on farnesylation targets the Lamins to the inner nuclear membrane is present in lamin A, B1 and B2 but absent in C. The localization of the lamin A/C isoform differs from the B isoform. The lamin A is also found in the nucleoplasm and the lamin found in the periphery, i.e., close to the nuclear envelope differs from the nucleoplasmic lamins. The nucleoplasmic lamin A is suggested to be less polymerized and possesses more solubility which accounts for the more dynamic nature of the lamin A isoform compared to the lamin B isoform which is immobile and restricted to the outer lamina¹⁰⁰. The difference in the farnesylation is one of the factors responsible for the localisation of the lamins. The B-type lamins contain a stable C-terminal farnesyl modification, which promotes its anchoring to the membrane. On the other hand, A-type lamin is transiently farnesylated and they are also found in the nuclear interior. although the real reason is unclear^{98,101}.

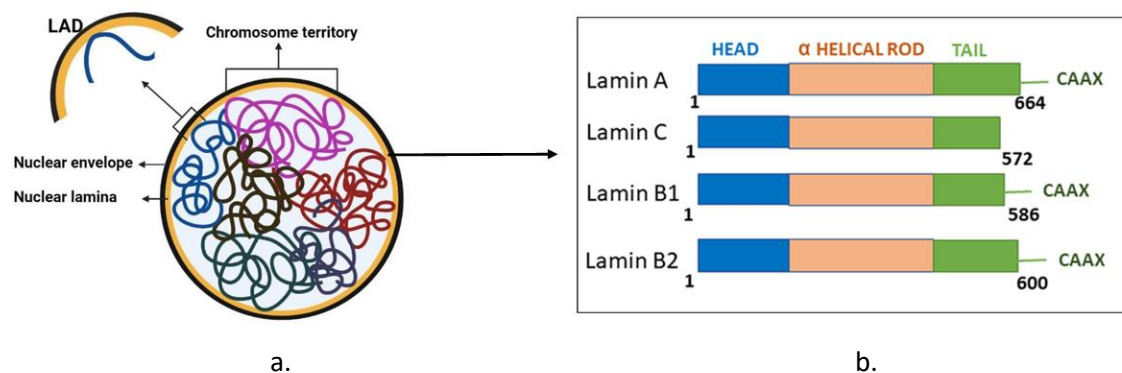


Figure 6.1: Schematic representation of the nuclear organization in eukaryotic cells and the structure of lamins. Panel a. is a schematic of the nuclear boundary and the lamina associated domains adapted from [94]. The different colour chromatin represents each chromosome territories. Panel b. is a representation of the general structure and domains in different types of lamins.

While it is known about the important sub-nuclear components responsible for the genome organization, it is essential to know what links the lamina and the chromatin. Because of its proximity to the chromatin, it is likely that the lamins would interact with the chromatin. LMNA mutations resulting in various laminopathies are known to alter the lamina and chromatin interaction and also alter the chromosome territories¹⁰². Various experimental techniques identify the regions of interaction between the proteins of interest and the chromatin. One compendious and far-reaching method of

establishing the protein-DNA interactions of the lamins at a genomic scale is the DNA adenine methyl transferase identification (DamID)^{3,103,104}. This technique exploits the possibility of adenine methylation in prokaryotes which is absent in eukaryotes. The principle behind DamID is the expression of a protein of interest (in this case lamin) to *E. Coli* Dam which interacts with chromatin. On being tethered to the protein of interest Dam transfers a methyl group to the GATC sequence in the closest proximity to the binding chromatin region. DpnI, the methylation-sensitive restriction endonuclease, fragments the genomic DNA in contact. The sequences can be amplified by PCR and detected by array-based or sequencing methods to generate a chromatin profile^{105,106}. Such genomic regions binding to the lamins are referred to as the lamina-associated domains (LADs). The LADs are cis-acting sequences that have a resolution of ~0.1-10Mb. This LAD signifies a specific interaction of the lamina with the genome and it is important to identify the nature of the genomic sequence associated with the LADs.

The discovery of the LADs and the binding protein partners of the lamins have expanded their potential role in the regulation of chromosome territories, positional gene silencing transcriptional regulation and other diverse functions directly or indirectly linked with the nuclear lamina. The LADs are reported to have a higher occurrence in the CTs that have a peripheral position⁹⁴. This work aims to understand in detail the key players in the LADs by identifying the sequence motifs. The signature motifs are at a resolution of 5-6 bases which is a much higher resolution than the reported experimental data. The characterisation of the LADs to identify the overrepresented motifs was done in different cell types in different organisms and the trend observed was consistent. The nature of the motifs obtained was AT-rich which is in agreement with the previously reported data that the constitutive LADs are AT-rich. To evaluate the evolutionary significance of these motifs, we have also repeated this analysis for different organisms, including mice and drosophila. By comparing the motif patterns across different organisms, we can identify motifs that are conserved across species and may play a conserved functional role in nuclear organization and gene regulation.

We used this information on the occurrence of different motifs of sizes 5 and 6 bps were used to train a hidden Markov model. The prediction of this mathematical model will help in distinguishing the LAD regions from the inter-LAD (iLAD). Identifying the

key player and a method of prediction of the likely lamin binding sites will be the prerequisites for understanding in detail the role of DNA protein interactions in chromosome positioning. This will give detailed insight into the molecular mechanisms behind laminopathies and other human diseases linked to lamina proteins and perhaps their possible cure. Overall, our approach provides a systematic and quantitative way to identify motifs that are enriched in the LAD regions and may have functional significance in regulating gene expression and nuclear organization. The analysis is highly scalable and can be applied to different organisms and different types of genomic data, making it a valuable tool for studying the role of nuclear organization in gene regulation and genome function.

6.2 Methodology

6.2.1 Demarcation of LAD and iLAD regions from experimental data:

To demarcate the LAD and the region interspersing the LAD (which we refer to as interLAD or iLAD), we establish a boundary between the two regions. We obtained the DamID data available under GEO Accession number: GSE69841¹⁰⁷. The contacts from the DamID data are at a resolution of 100kbps. For every 100kbps bin the data stores the single cell-wise contacts (for 118 cells) with the nuclear lamina. We looked at the LAD coverage at different contact frequencies. Contact frequency is defined as the proportion of the 118 cells that show a positive contact score in a 100kb bin.

$$\text{Contact Frequency (CF)} = \frac{\text{Number of cells with positive score in DamID data}}{118}$$

$$CF \text{ cutoff} = \frac{118 - CF}{118}$$

We demarcated the chromosome into LAD and iLAD regions based on the CFs in each 100kbps bin. Those regions in the chromosome where the contact frequency is above the cutoff are defined as LADs and the rest of the chromosome is defined as iLADs.

6.2.2 Mapping motifs in the chromosome and comparing distribution in LAD vs iLAD regions:

We positioned the occurrences of all possible combinations of 5 or 6mer motifs by brute force defining them as the motif vector. The motif vector of motifs of length k will have 4^k combinations because of 4 possible bases A, T, G, C. The position index of all the individual motifs in the motif vector was obtained using a pattern searching algorithm (naive string search) and stored in a .json file (dictionary with key as motif and list of positions as values) for each chromosome. To obtain the abundances of the motif in the LAD to iLAD regions we used a metric called the LAD to interLAD ratio (LIR). LIR is the normalized count of motifs in LAD to the interLAD region. The normalization was done according to the LAD coverage at the given contact frequency cutoff.

$$LAD\ to\ iLAD\ ratio(LIR) = \frac{Normalized\ motif\ count\ in\ LAD}{Normalized\ motif\ count\ in\ iLAD}$$

$$Normalized\ count = \frac{Count\ in\ LAD/iLAD}{LAD/iLAD\ length\ in\ Mbps}$$

The LIR was used to get descriptive statistics of motifs for sizes 5-6 across the different chromosomes in the organism of interest.

6.2.3 HMM- Viterbi prediction of LADs and iLADs:

We designed a two-state Hidden Markov model to predict and distinguish a LAD from an interLAD region given a genomic sequence of interest. To find the most likely sequence of hidden states, we employed the Viterbi algorithm, which is a dynamic programming algorithm used to find the optimal path of hidden states given a sequence of observations. The algorithm works by recursively computing the probability of each possible next hidden state, based on the probability of the current hidden state and the transition probabilities between hidden states. The algorithm generates a matrix of probabilities for each set of observations in a hidden state. We then select the hidden state with the highest probability and use it to update the most likely sequence of LAD and iLAD states.

In our approach, we used the observation sequences obtained from the window-wise reading frame of the genome, with the window size corresponding to motif sizes 5 and 6. To train the model, we randomly selected 7 chromosomes (chromosomes 1, 3, 9, 12, 15, 17, 18) from the human chromosome dataset to create a training set, ensuring that there was no bias towards a LAD-enriched or iLAD-enriched chromosome. The starting probability is considered to be 0.5 (equal probability of selection of LAD or iLAD states) The Viterbi predictions were calculated using the following parameters:

$$\begin{aligned}
& \text{States} : \{LAD, iLAD\} \\
& \text{Observation window} : \{Motif_1, Motif_2, Motif_3 \dots \dots Motif_{N-k+1}\} \\
& \text{Motif size} : k \\
& \text{Length of sequence} : N \\
& \text{Number of reading windows} : N - k + 1 \\
& \text{Transition Probability (TP)} : \begin{cases} TP_{LAD \rightarrow iLAD} \\ TP_{iLAD \rightarrow LAD} \\ TP_{LAD/iLAD \leftrightarrow LAD/iLAD} \end{cases} \\
& \text{Emission probability (EP)} : \begin{cases} Motif_{LAD} \\ Motif_{iLAD} \end{cases} \\
& \text{Starting probability (SP)} : \begin{cases} LAD: 0.5 \\ iLAD: 0.5 \end{cases}
\end{aligned}$$

Since the experimental dataset was at a resolution of 100kbps and the predictions were made at a 10^5 order higher resolution of moving windows, we converted the predictions made at the resolution of moving windows to a 100kbps bin using a cutoff (n). In a 100kbps stretch, if at least n moving windows have a prediction of LAD states then the entire 100kbps stretch will be considered as a LAD and vice versa.

6.2.4 Identifying protein binding partner(s) for the LAD enriched motifs:

A dataset of all the interactors of Lamin B1 was obtained using String database¹⁰⁸. The shortlisting of the candidates for the dataset was done using experimental data, co-expression patterns and literature-based text mining. The results were further shortlisted to ones which have a structure available in the PDB. In case of unavailability of their structures from PDB the Alpha fold^{109,110} models were used.

JEDII (Juxtaposition Enabled DNA interface Identifier) was used to superimpose the query proteins with the optimal protein-DNA structure from the database. JEDII is a template-dependant algorithm that studies the DNA–protein interface to identify hydrogen bond donor-acceptor pattern similarity with the database of template structures². The information of the hydrogen bond donor-acceptor atoms to align the query and the template is done using CLICK^{8,111}. The best model is selected based on the query overlap percentage, calculated as the number of donor and acceptor matches between the query and template proteins relative to the total number of donors and acceptors in the query protein. DNA-binding residues are identified as amino acids within 4 Å of any DNA atom, and the DNA sequence from the template is predicted as the binding sequence for the query protein. This methodology, named Juxtaposition DNA-binding Interface Identifier (JEDII), has been validated using the PDNA-285 dataset and other benchmark datasets. DNAProDB^{112,113} was used to extract the features of interactions of the DNA consensus from the best-fit template with the query protein. The hydrogen bond information of the query protein with the consensus was used to compare with the shortlisted motifs based on LAD-iLAD distribution.

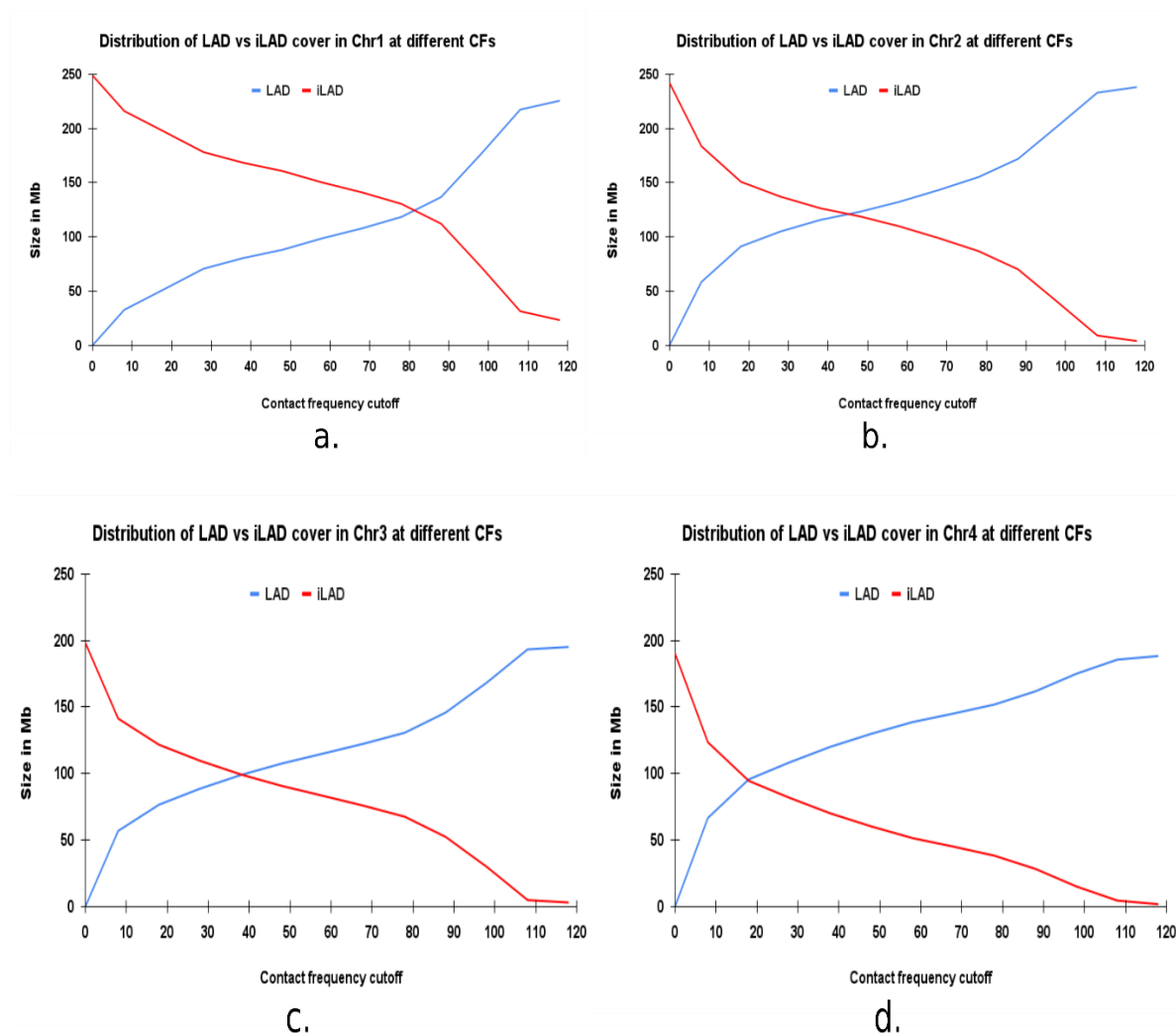
6.3 Results

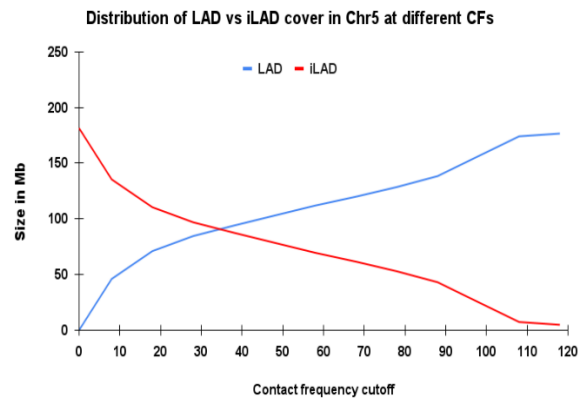
6.3.1. Distribution of LAD-iLAD regions across chromosomes:

From the DamID data, we extracted the contact frequency for the 100 kbps bins and obtained the LAD coverage at different thresholds of contact frequency. At different CF cutoffs, we plotted the distribution of the LAD coverage of the chromosome ([figure 6.2](#)). From the ranges of LAD coverage of different chromosomes at different thresholds, we established a CF cutoff of ~60%. This signifies that only those 100kb bins from a chromosome will be considered as LADs where at least 60% of the 118 cells establish contact with the nuclear lamina. The cutoff was chosen such that we are neither too stringent that we miss on including the chromosomes having smaller LAD coverage nor too lenient that the regions having low confidence of being LADs are also considered the same. With the established cutoff we obtain the LAD coverage across different chromosomes as shown. It is to be noted that chromosomes 8, and Y are missing from the cell line used in the DamID dataset. We also failed to include

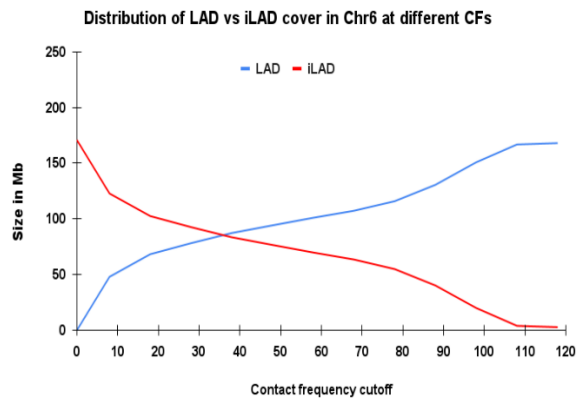
chromosome 22 in our dataset with the given cutoff as no 100kbps genome stretch established contact above the contact frequency. At the established cutoff the chromosome with the highest and the smallest LAD coverage is chromosome 4 and 19 respectively ([figure 6.3](#)).

We also wanted to look at the gene density of different chromosomes at the established LAD coverage if there is a direct correlation. The gene density is calculated as the number of genes per Mbps of a chromosome. Chromosome 19 has the highest gene density (~22 genes per Mbps) and chromosome 18 has the lowest (~3 genes per Mbps)⁶⁸. Our analysis shows that at the established LAD and iLAD boundaries in different chromosomes there is a direct negative correlation with gene density (Pearson correlation coefficient -0.68) ([figure 6.4](#)). That is chromosomes with high LAD coverage have low gene density and vice versa. This is in agreement with the literature that LAD regions do not overlap with genic regions.

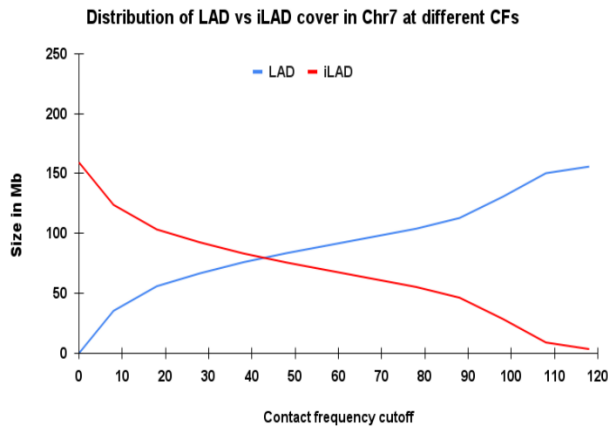




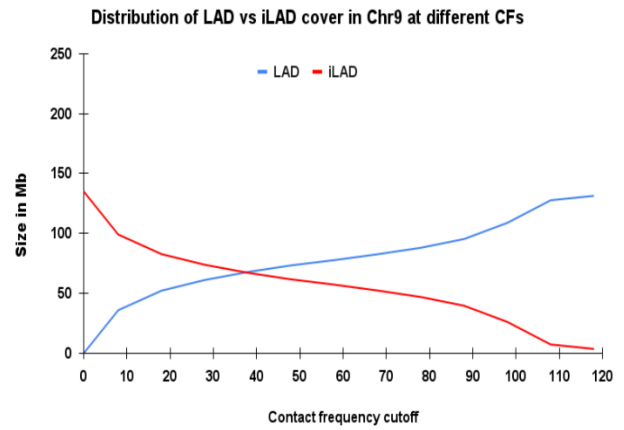
e.



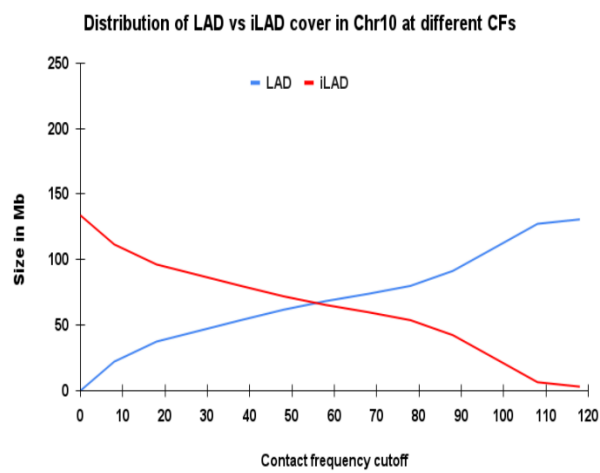
f.



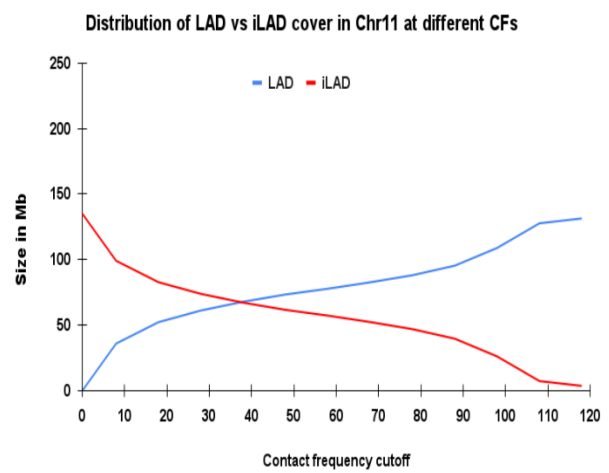
g.



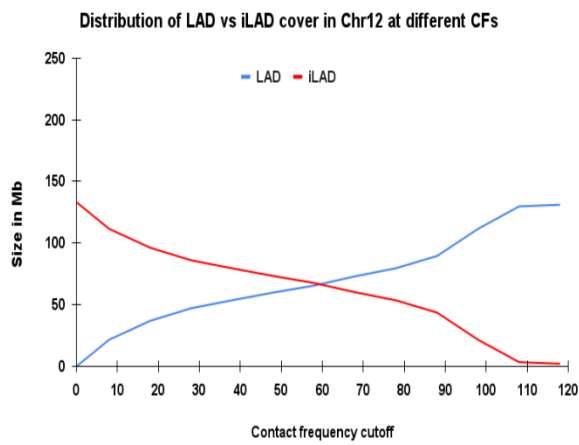
h.



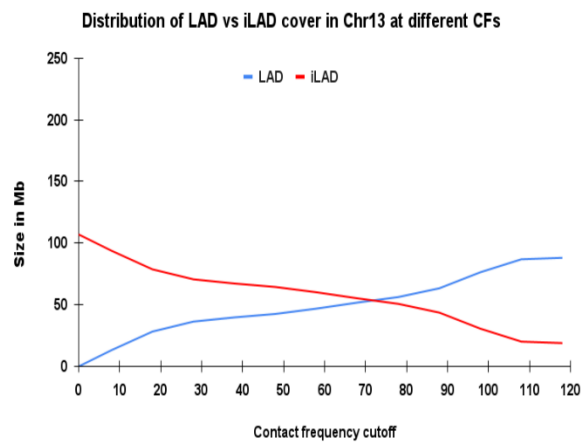
i.



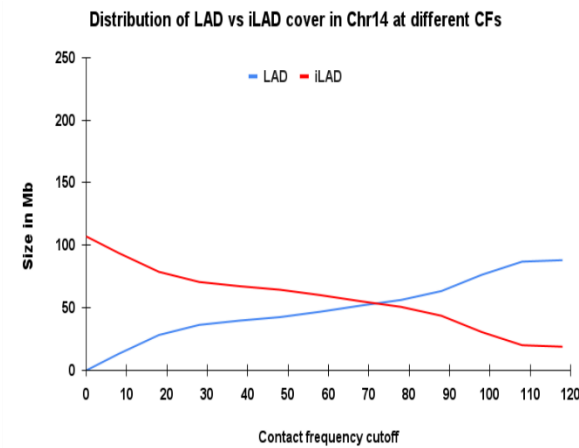
j.



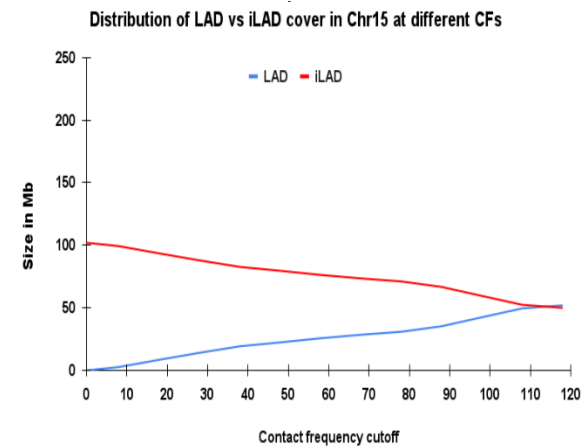
k.



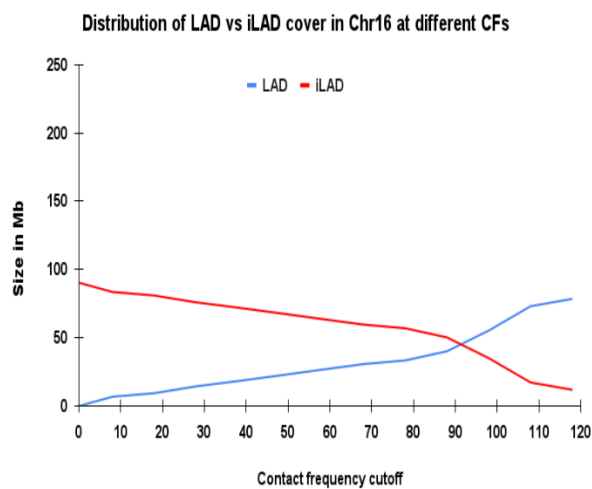
l.



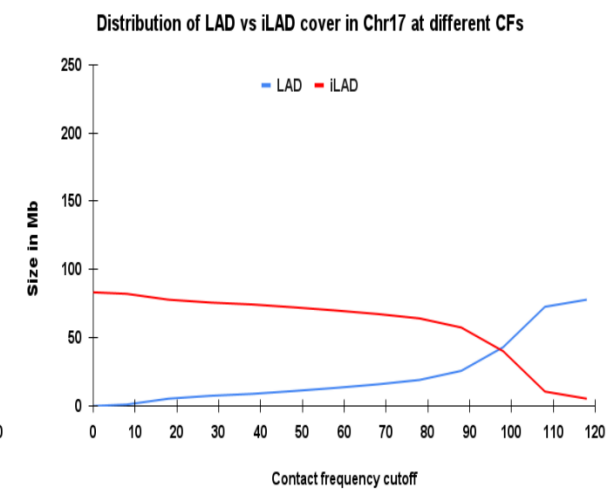
m.



n.



o.



p.

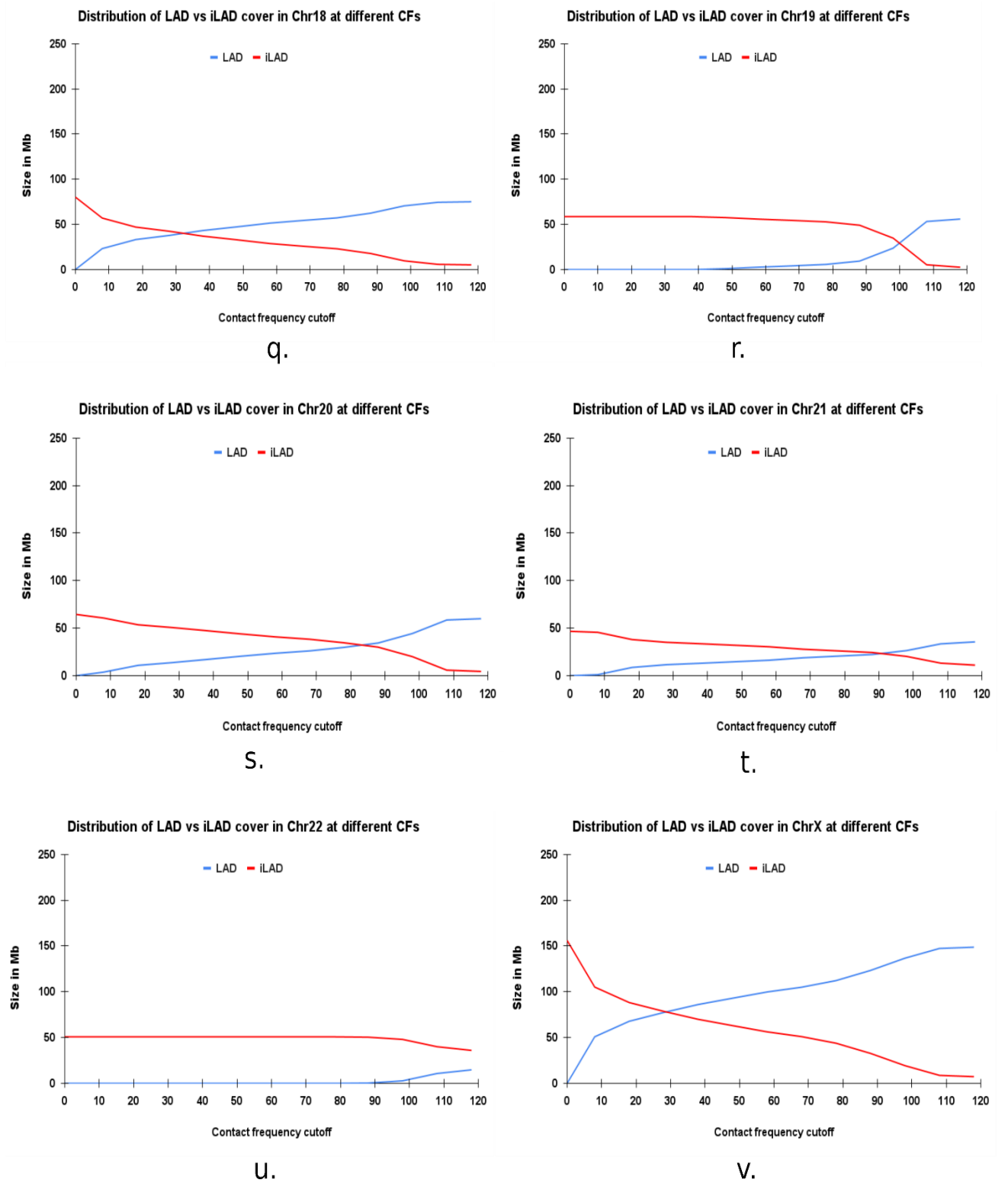


Figure 6.2: LAD vs iLAD coverage at different contact frequency cutoffs. The x-axis represents the number of cells which show a contact according to DamID data. The y-axis represents the chromosome cover in Mbps. The blue and the red lines represent the coverage for LAD vs iLAD respectively. Legends 1a to 1v show the plots for chromosome 1 to X.

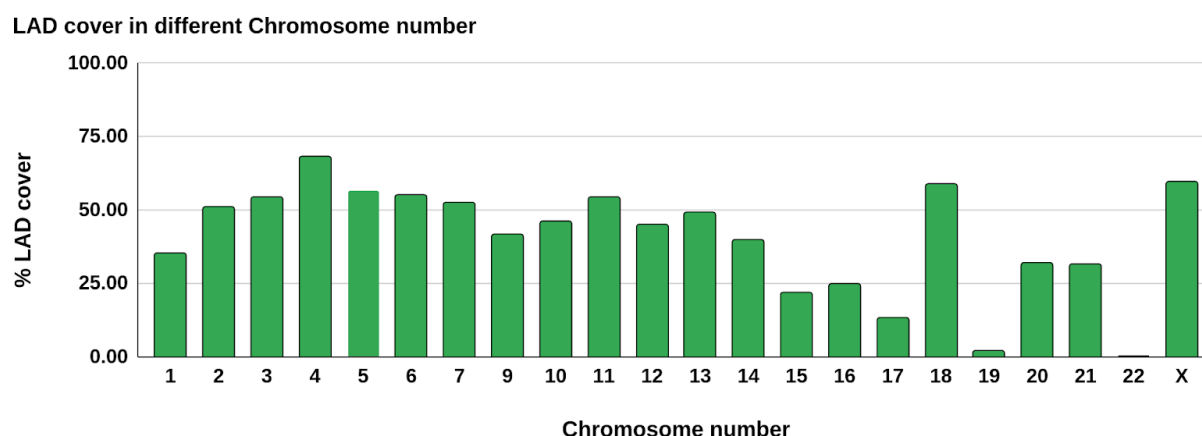


Figure 6.3: Distribution of LAD coverage across different chromosomes at the contact frequency threshold of 60%. The x-axis represents the chromosome number and the y-axis represents the % LAD coverage for the individual chromosomes. The green colour bars represent the %LAD coverage for individual chromosomes.

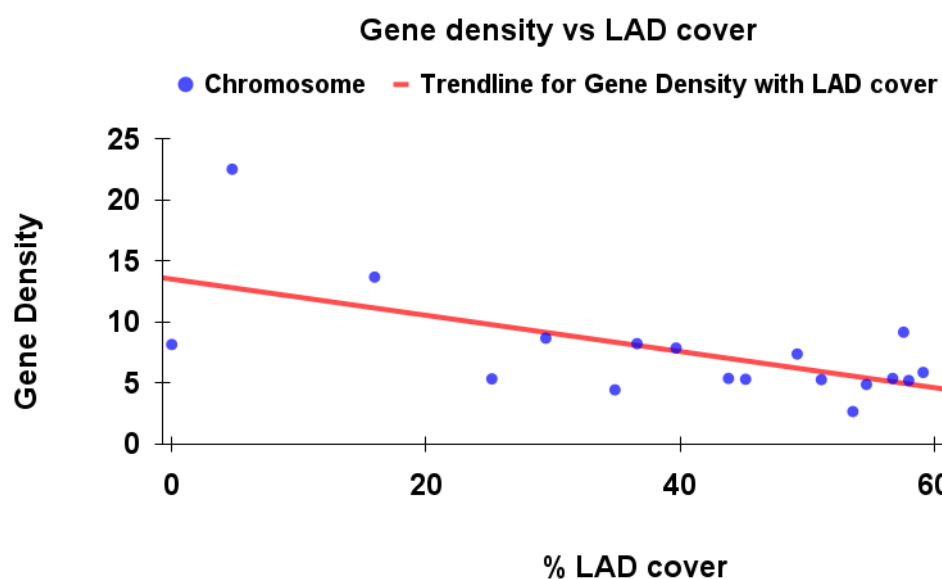
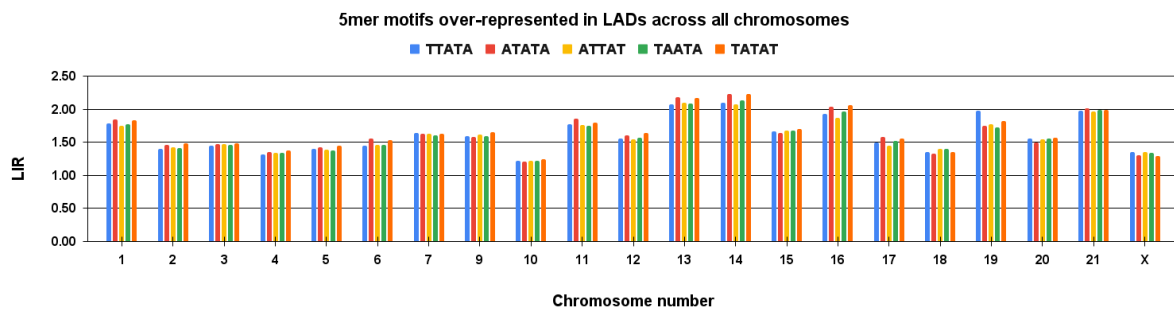


Figure 6.4: Comparison of LAD coverage across different chromosomes with gene density. The x-axis represents the LAD coverage and the y-axis represents the gene density across different chromosomes.

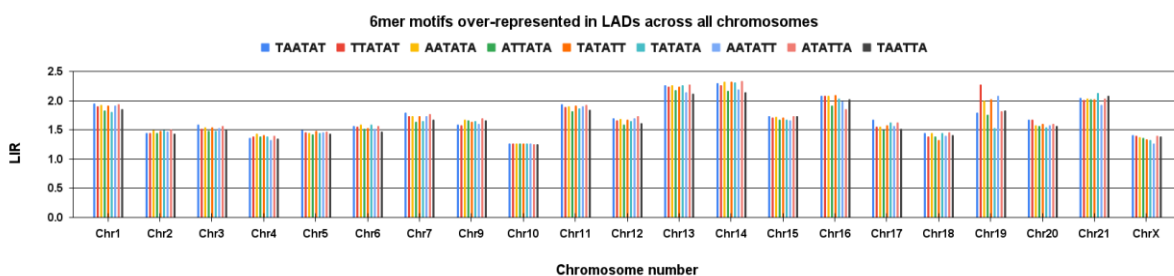
6.3.2. Identification of motifs enriched in LAD and iLAD regions:

We looked at the abundance of motifs of size 5 and 6 mers in the LAD and the iLAD regions using the LAD to iLAD ratio(LIR). The metric takes into consideration the normalized abundances of a motif in the LAD region compared to iLAD regions. An LIR greater than 1 indicates higher occurrences in the LAD regions. Similarly, LIR less than 1 indicates an abundance of the motif in the iLAD regions. We calculated the motif abundances for all combinations of motifs of sizes 5 and 6 across different chromosomes. Using a LIR cutoff of 1.2, We identified five 5mers (TTATA, ATATA, ATTAT, TAATA and TATAT) that were overrepresented in the LAD regions compared to the iLAD regions ([Figure 6.5a](#)). Similarly, the nine 6mers were identified (TAATAT, TTATAT, AATATA, ATTATA, TATATT, TATATA, AATATT, ATATTA, and TAATTA) which had LIR greater than 1.2 across all chromosomes ([Figure 6.5b](#)). Using a similar estimate of cutoff we identified three 5mers (CGGGA, ACGGG and CGCCG) and five 6mers (CGTTGG, GGAGCG, AACGGG, TCGTTG and CGGACG) which had LIR lesser than 0.8, i.e. they were overabundant in the iLAD regions ([Figure 6.5c and 6.5d](#)). We also that the motifs overrepresented in the 5meric forms are a subset of the 6mers abundant in both LADs and iLADs.

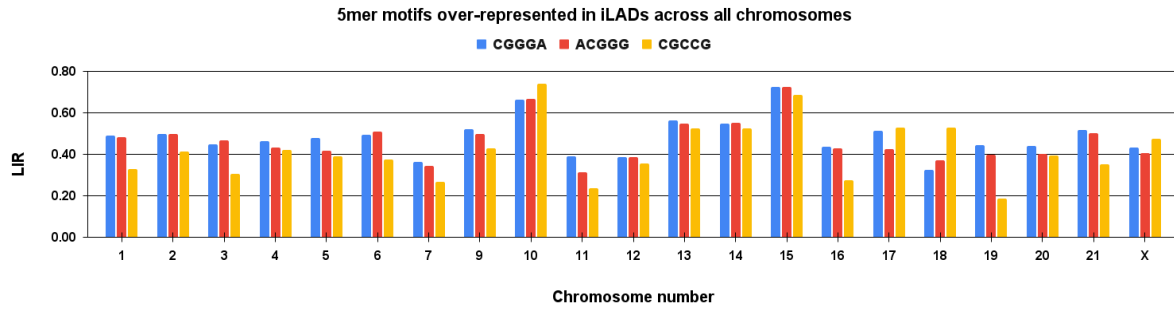
5a.



5b.



5c.



5d.

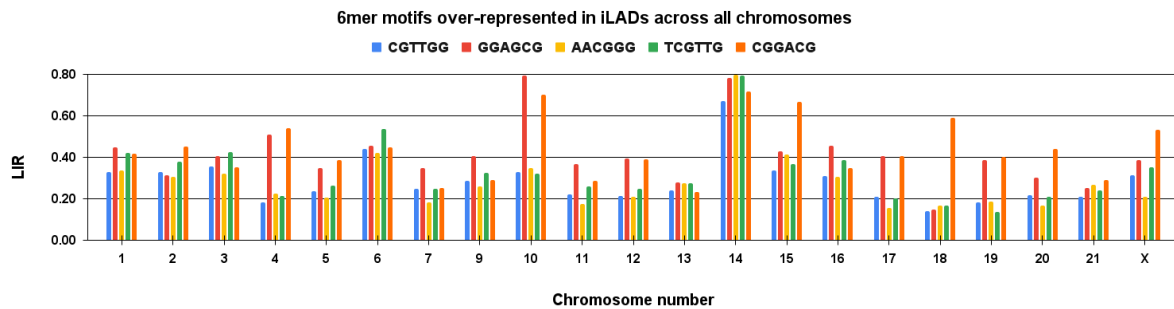
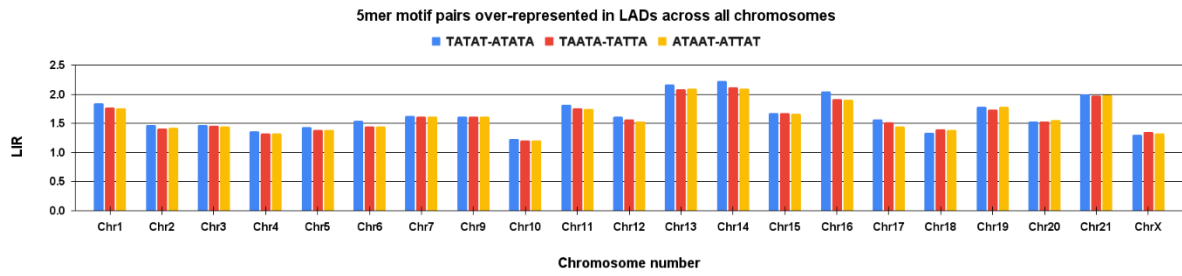


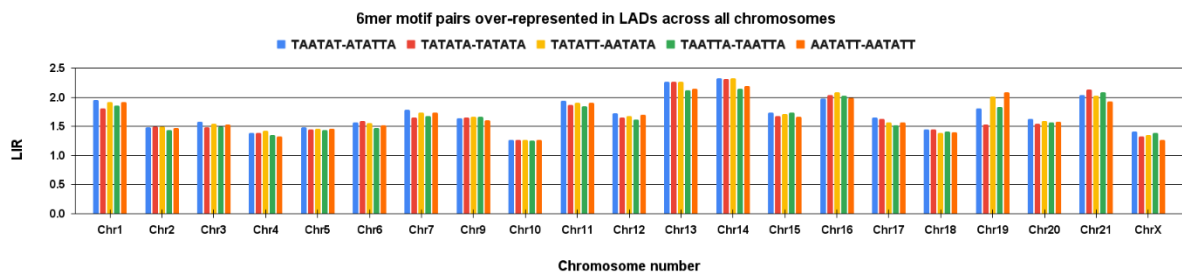
Figure 6.5: LIR of 5mer and 6mer motifs over-represented in LADs and iLADs across different chromosomes. The y-axis represents the LIR and the x-axis represents the chromosome number. Figures 5a. and 5b. represents the LIRs of motifs enriched in LAD regions in 5 and 6mers respectively. Figures 5c. and 5d. represents the LIRs of motifs enriched in iLAD regions in 5 and 6mers respectively.

We additionally also looked at motif abundances using both forward and reverse complements. We used the average LIR of motifs and their reverse complement and then observed the abundance of motifs. We identified three pairs of 5mer motifs (TATAT-ATATA, TAATA-TATTA and ATAAT-ATTAT) having higher abundance in LAD compared to iLAD across all chromosomes ([figure 6.6a](#)). Similarly, five pairs of 6mer motifs (TAATAT-ATATTA, TATATA-TATATA, TATATT-AATATA, TAATTA-TAATTA and AATATT-AATATT) in the forward and reverse complement representation were identified to be enriched in the LAD regions ([figure 6.6b](#)). The 5mer motifs enriched (i.e. having LIR greater than 1.2) are a subset of the 6mer motif pairs and also overlap with the singular motif representation. This confirms the conservation of the motif pattern using the LIRs. Similarly, six motif pairs in both 5 and 6meric representations were identified to be overrepresented in the iLAD regions, with all the motifs having a consistent LIR lesser than 0.8 across all the chromosomes ([figure 6.6c. and 6.6d.](#)).

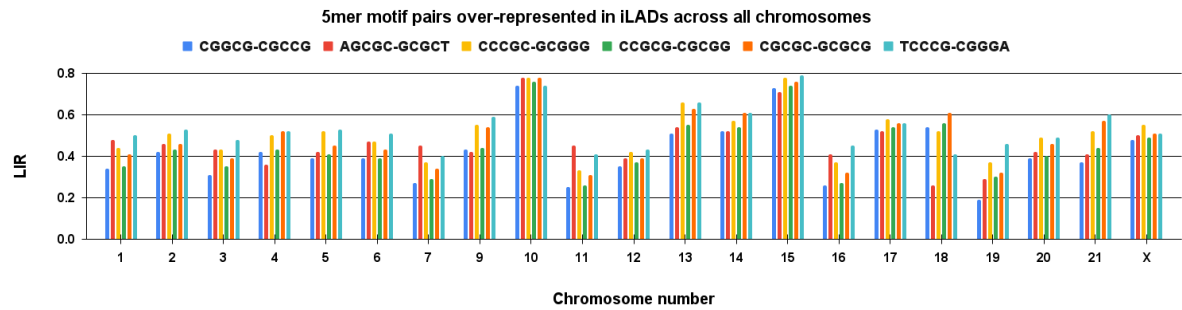
6a.



6b.



6c.



6d.

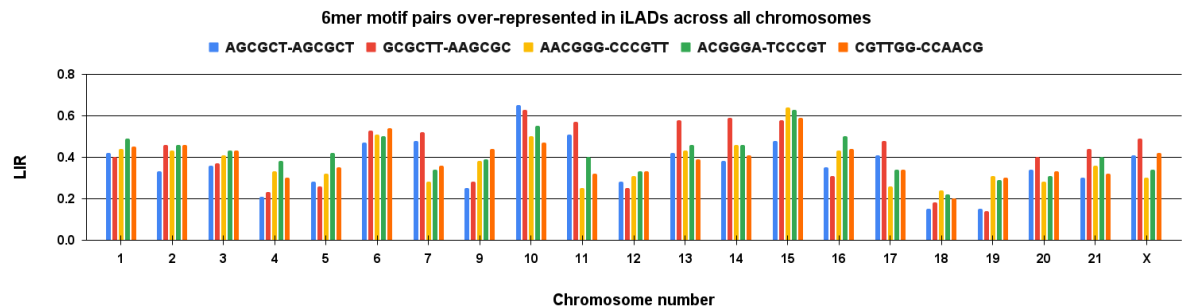
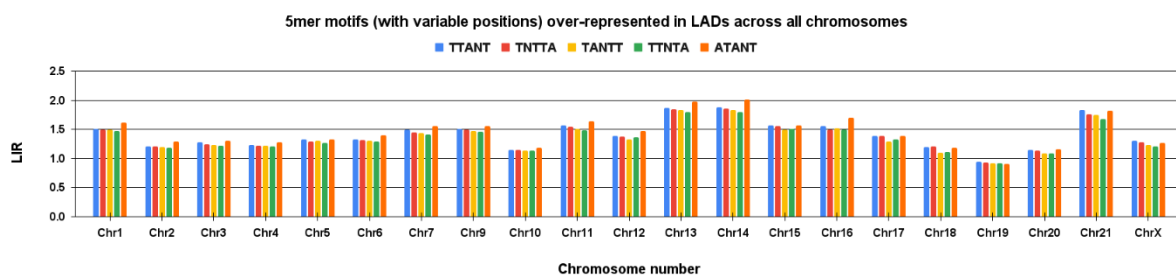


Figure 6.6: LIR of 5mer and 6mer motif pairs (forward and reverse complement) over-represented in LADs and iLADs across different chromosomes. The y-axis represents the LIR and the x-axis represents the chromosome number. Figure 6a. and 6b. represents the LIRs of motifs

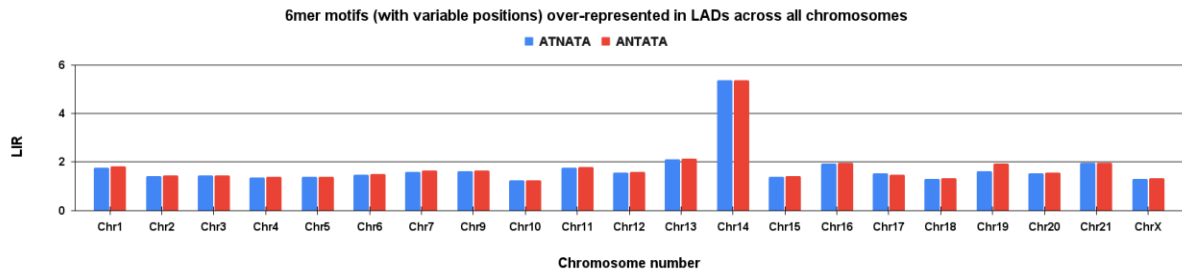
enriched in LAD regions in 5 and 6mers respectively. Figures 6c. and 6d. represents the LIRs of motifs enriched in iLAD regions in 5 and 6mers respectively.

We also looked at motif representations at 5 and 6meric forms introducing variable positions in the motif representations. For a more generalized representation of motifs, we computed the LIRs using a variable base (N) at different positions in the motifs and will be referring to them as N-mers. For example, motifs: ATTAT, AGTAT, ACTAT and AATAT in the variable representation will be represented as ANTAT. We repeated the analysis introducing N at positions in the motifs except the first and the last position as the bases in these positions represent the flanking sites to the specific recognition sequences. We identified five 5mers (TTANT, TNTTA, TANTTA, TTNTA and ATANT) having LIR greater than 1.2 across all chromosomes except chromosome 19 ([figure 6.7a](#)). For chromosome 19 we selected the motifs which have the highest LIRs in this N-meric representation. Similarly, for 6 mers we identified two motifs (ATNATA, and ANTATA) having LIR greater than 1.2 across all chromosomes thereby having higher abundance in the LAD regions compared to the iLAD ([figure 6.7b](#)). Similarly, we identified three 5meric motifs (GNGCG, CGNCG and CGCCG) all having LIR lesser than 0.8 across all chromosomes thus having higher abundance in the iLAD regions ([figure 6.7c](#)). For 6mers we identified three motifs (CGTNGG, CGNCGG and ANCGGG) which had LIR lesser than 0.8 across all chromosomes except chromosome 14 ([figure 6.7d](#)). It is to be noted in the N-mer representations these were the three N-meric motifs which had the lowest LIR in chromosome 14.

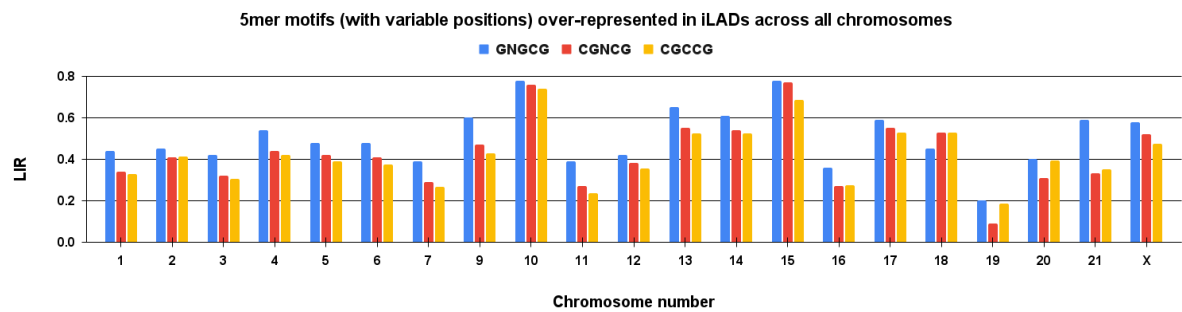
7a.



7b.



7c.



7d.

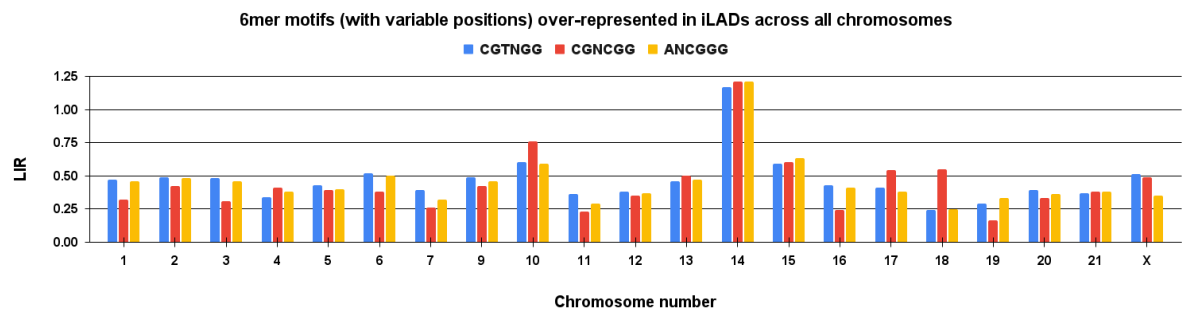


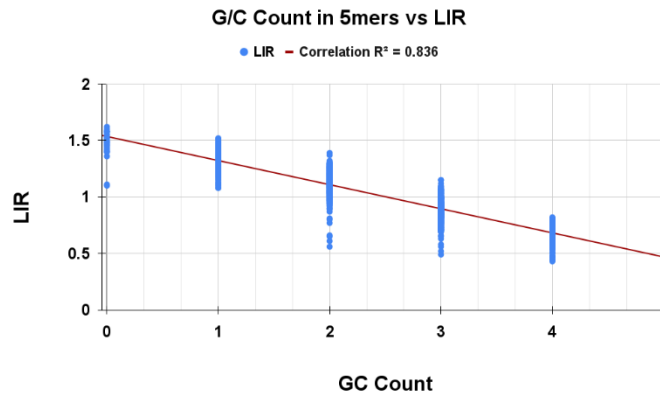
Figure 6.7: LIR of 5mer and 6mer motifs with variable positions (N-mers) over-represented in LADs and iLADs across different chromosomes. The y-axis represents the LIR and the x-axis represents the chromosome number. Figure 7a. and 7b. represents the LIRs of motifs enriched in LAD regions in 5 and 6mers respectively. Figures 7c. and 7d. represents the LIRs of motifs enriched in iLAD regions in 5 and 6mers respectively.

From the LIRs across all different representations, we saw the motif pattern being conserved and the 5meric motifs are a subset of the 6meric motifs which are overrepresented in the LAD as well as the iLAD regions.

6.3.3. Motifs enriched in LADs are AT-rich

The conserved 5 and 6meric motifs overrepresented in LAD and iLAD regions showed a preference for the AT-rich motifs in all the different representations. Similarly, the motifs overrepresented in the iLAD regions showed a preference for GC richness. We compared the GC count of a motif with the LIR to understand if there is a conserved trend. We observed that there is a strong negative correlation between the GC count in a motif and the LIRs. That is motifs which have higher LIRs (or are overrepresented in the LAD regions) have less GC count (and vice versa). The trend is consistent for 6mers. Both 5 and 6 mer motifs enriched in LAD's low preference for GC richness (PCC is -0.91 and -0.89) respectively ([figure 6.8a and 6.8b](#)). This is in agreement with previously reported data that the constitutive LAD regions have a higher occurrence of AT-rich isochores. Our observation of the motif abundances is in agreement with the reported data.

8a.



8b.

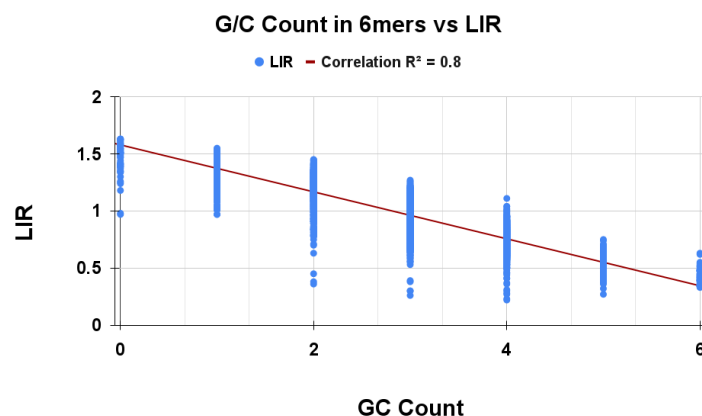


Figure 6. 8: Scatterplot of comparison GC count in motifs vs the LIRs. The y-axis represents the LIR and the x-axis represents the GC count of the motifs. The trendline is shown in red and the blue colour represents each motif data point. Figures 8a and 8b show the scatterplot for 5 and 6mer respectively.

6.3.4. Identification of evolutionarily conserved motifs enriched in LAD and iLAD regions:

We obtained a similar interaction dataset of lamina and chromatin interactions for other organisms to see if there is any conservation of the pattern of motifs overrepresented in the LAD vs iLAD regions. We obtained the DamID dataset for three different cell types in mouse¹¹⁴ and *Drosophila*¹¹⁵. Both the DamID datasets obtained captured the interaction of different chromatin segments with lamin B in mouse and LAM (only B-type lamin) in *Drosophila*. It is to be noted that the resolution of the chromatin contacts for both mouse and *Drosophila* were of a lower resolution (0.25~0.5Mbps for mouse and 0.5~1.5Mbps for *Drosophila*) compared to the human DamID dataset in our study (~0.1Mbps). We calculated the LIRs for three cell types in mouse (mouse embryonic fibroblasts (MEF), embryonic stem cells (ESC) and mouse astrocytes) and *Drosophila* Kc cells. All the different cell types in mice show a conserved overlap of the LIR ratios for the constative motifs, thereby confirming the conservation of motif pattern and identification of the constitutive LADs or the cLADs¹¹⁶, which remains conserved independent of the cell type. The LIRs in mouse for all the motifs follow the same trend compared to its human counterparts ([figure 6.9](#)). The analysis suggests that enrichment of the motifs in LADs are conserved evolutionarily. The list of evolutionary

conserved motifs which have more occurrence in the LAD regions compared to the iLAD regions across all organisms and cell types can be found in [table 6.1](#).

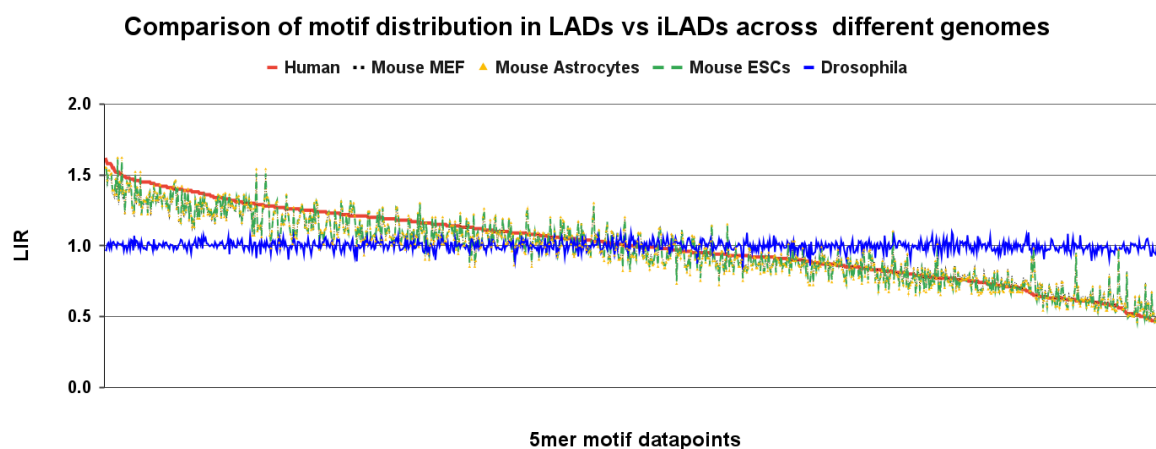


Figure 6.9: Distribution of LIR for different 5mers across different species. The red, black, yellow, green and blue lines represent the LIRs for human, mouse MEF cells, mouse astrocytes, mouse ESCs and Drosophila respectively. The y-axis represents the LIRs and the x-axis represents the 1024 5mer motif data points.

Table 6.1: List of evolutionary conserved 5 and 6mer motifs with an overabundance in LAD regions (high LIRs) across different species and cell lines.

Conserved 5mer motifs	Conserved 6mer motifs
AATAA	AATTTA
AATTA	ATAATA
ATAAT	ATATAT
ATATA	ATTAAT
ATTAA	ATTATA
ATTAT	ATTATT
TAATA	TATATA
TAATT	TATTAT
TATAT	
TATTA	
TTAAT	
TTATT	
TATTT	
TATTT	
TCTAT	
TTTTA	
ATTTT	
AAAAT	

6.3.5. Prediction of LAD - iLAD region using Hidden Markov model:

We used the LIRs of all 5 and 6 meric motifs to train a Viterbi implementation of the Hidden Markov model to predict the LAD and the iLAD states given a genome sequence. Our model reads the input genome sequence in windows of size 5 and 6mers and predicts two resolutions: a) Prediction based on each moving window and b) Prediction at 10kbps and 100kbps stretches. The experimental results are of the resolution 100kbps.

For moving window-wise prediction, we read the genome in window size corresponding to the motif size. We obtain the emission probabilities(EP) and the transition probabilities (TP) from the 7 chromosomes in the training dataset. We selected 7 chromosomes randomly to train the HMM. The 7 chromosomes in the training dataset are chr1, chr3, chr9, chr12, chr15, chr17 and chr18. The remaining chromosomes were included in the testing dataset. The average MCC (training and testing set combined) for the window-wise prediction is around ~0.2 ([table 6.2](#)) for both motif sizes 5 and 6. It is to be noted that the results for the experimental data for comparing our prediction results are in the order of 100kbps which is 10^5 times lower than our prediction. We converted the experimental data (captured by DamID) for 100kbps LAD bins to window-wise resolution.

As a control, we compared the performance of the model using the EP of individual bases (ATGC). The TP remains unchanged. The model underperforms compared to the motif window-wise prediction in both terms of MCC and accuracy ([table 6.2](#)). Thus, just the base frequency is not sufficient for the determination of the LAD or the iLAD orientation, rather the specific combination of bases in terms of motifs is a better metric of prediction.

Table 6.2: HMM performance for moving windows at different resolution

Resolution	MCC	Accuracy
Moving window A/T/G/C	0.19	63.05%
Moving window 5mer	0.20	63.26%
Moving window 6mer	0.22	63.01%
10kb bin wise prediction (using 5mers)	0.28	68.51%
100 kb bin wise prediction(using 5mers)	0.36	73.73%
10kb bin wise prediction (using 6mers)	0.29	69.05%
100 kb bin wise prediction(using 6mers)	0.38	74.19%

The second method of prediction at the same resolution as that of the experimental data had an additional step. We converted our window-wise prediction to 10 and 100kbps resolution. We combined the prediction for each position to 10 or 100kbps bins and established a cutoff to categorize the entire bin as LAD or iLAD. We established an optimum cutoff of 70% i.e. at least 70% or more of the positions in the bin of interest have a prediction of being a LAD to define the entire bin as LAD. We observed that the prediction model performs better both in terms of MCC and accuracy for both 5 and 6-meter motif resolutions. The model performs best at 100kbps resolution (same as that of the experimental predictions) when reading the genome at a window size of 6meric motifs with MCC and accuracy of 0.38 and 74.19% respectively.

Additionally, we also tested the performance of the model using 100 shortlisted motifs which are conserved in the LAD and the iLAD regions (50 LAD enriched and 50 iLAD enriched motifs). We observed that the model performed better with respect to all metrics ([table 6.3](#)). These motifs can be further studied in identifying the protein binding targets of the LADs (or iLADs)

Table 6.3: Summary of all metrics for HMM prediction of LAD-iLAD states

	Sensitivity	Specificity	FPR	FNR	MCC	Accuracy	Precision
All motifs (5mer)	0.74	0.61	0.39	0.26	0.36	0.74	0.59
Shortlisted motifs (5mer)	0.6	0.83	0.17	0.4	0.44	0.73	0.77

6.3.6. Identifying protein targets associated with LADs:

Using the STRING database, we obtained 32 proteins that have reported protein-protein interaction with Lamin B1. The structures of the shortlisted proteins were obtained from PDB and Alpha fold. The 8 and 14 of the shortlisted proteins had a localization reported in the nuclear envelope and the nucleus respectively. The other proteins have reported localization in the cytoplasm endoplasmic reticulum and the extracellular membrane. The JEDII was used to find the template structure with there is best overlap with the query proteins. The consensus was obtained and compared to our shortlisted motifs. 13 out of 32 of the query proteins showed complete overlap with our shortlisted motifs. The motifs of 5 and 6 mers were obtained from the consensus by looking at the hydrogen-bound bases of the template DNA with the query protein ([table 6.4](#)). These proteins are likely to be the targets for specific interaction with the abundant motifs in the LADs.

Table 6.4: Shortlisted protein targets that are interacting with Lamin B1 based on LAD. The proteins highlighted in red are the ones which show overlap with shortlisted motifs.

Serial Number	Query Protein	Query PDI	Template PDE	Structural Overlap	RMSD (Å)	Consensus
1	Human Barrier to Autointegration Factor (BAF)	1ci4	2bzf	100%	0.76	GTGGAGG
2	Human Cyclin Dependant Kinase 2	1e1x	1jko	93.48%	2	TGTTTTTGGTAAGA
3	Importin Subunit Beta-1	1f59	1jj6	89.36%	2.21	ATCTTATCAAAAAC
4	Granzyme B	1iau	4dii	93.01%	1.31	GGTTGGTGTGGTTGG
5	Catenin beta-1	1jdh	1jko	95.65%	2.09	TGTTTTTGGTAAGA
6	Caspase-3	1nms	1tc3	94.12%	2.22	AGGGGGGGTCTATAGAACTT
7	Caspase-6	2wdp	1jko	93.48%	2.01	TGTTTTTGGTAAGA
8	Caspase-8	1qtn	1tc3	94.12%	2.2	AGGGGGGGTCTATAGAACTT
9	Caspase-9	1nw9	1hcr	92.31%	2.18	TGTTTTTGATAAGA
10	LIM domain only protein 7	2eaq	3u6p	74.16%	2.23	GGTAGATCCCGAC
11	FH1/FH2 domain-containing protein 1	3dad	1jkq	95.65%	2.05	TGTTTTTTATAAGA
12	Chromobox protein homolog 3	3kup	1wd0	80.30%	1.91	CCTATATAGG
13	CAAX prenyl protease 1 homolog	5syt	1jko	85.11%	2	TGTTTTTGATAAGA
14	Cyclin-dependent kinase 2	5uq2	1jj6	93.62%	2.06	ATCTTATCAAAAAC
15	Telomeric repeat-binding factor 1	5wir	1jj8	93.88%	2.19	TTTTGATAAG
16	LEM Domain of Emerin	6ghd	2bzf	100%	0.56	GTGGAGG
17	MBD domain of Methyl-CpG-binding protein 2	6ogj	5bt2	100%	1.08	ATAGAAGAATTCGTTCCAG
18	Histone-lysine N-methyltransferase setd3	6ox3	1ig7	89.66%	2.06	CACTAATTGAAGG
19	Lamin binding SUN domain-containing protein 1	4dxr	1wtq	85.94%	2.23	ATCTTATCAAAAAC
20	Lamin binding SUN domain-containing protein 2	4dxt	1wd1	85.94%	2.09	CTACGTAGG
21	Bromo1 domain of Bromodomain-containing protein 4	7rjn	1jkr	91.30%	2	TGTTTTTGACAAGA
22	Bromodomain-containing protein 4	7rjr	1jko	89.13%	2.05	TGTTTTTGGTAAGA
23	Guanylate Cyclase	Alpha Fold	1tc3	84.31%	2.01	AGGGGGGGTCTATAGAACTT
24	Guanylin	Alpha Fold	1jjw	80.85%	1.97	TGTTTTTGATAAGA
25	Heat-stable enterotoxin receptor	Alpha Fold	1jkr	91.30%	2.05	TGTTTTTGACAAGA
26	Lamina associated polypeptide 2 isoform alpha	Alpha Fold	1nwq	58.33%	1.29	TTCCTATTGCGCAATCCAGTT
27	Lathosterol oxidase	Alpha Fold	1jko	89.13%	1.94	TGTTTTTGGTAAGA
28	LEM domain containing protein 2	Alpha Fold	1tc3	88.24	2.04	AGGGGGGGTCTATAGAACTT
29	Methylsterol Monooxygenase	Alpha Fold	5lxu	94.55%	2.28	ATCGTATCT
30	Proline rich protein 14	Alpha Fold	5d23	55.74%	1.8	GAATCTATGTAAA
31	ZW10 interactor	Alpha Fold	1nwq	55.83%	1.27	TTCCTATTGCGCAATCCAGTT
32	Histone deacetylase 3	4a69	1jj6	93.62%	2.06	ATCTTATCAAAAAC

6.4 Discussion

In this study, we attempted to characterise the LAD and the iLAD regions of the human genome at a resolution of 5 and 6 meric motifs at a resolution where specific interaction with protein partners occurs. We used the 100kbps resolution of the DamID data of the genome and LaminB1 interactions¹⁰⁶ to study the sequence of the LADs vs the iLADs. We established and studied the LAD coverage at different contact frequencies and used a suitable cutoff (~60%) to generate our dataset of LAD and the iLAD regions across different chromosomes. We observed a negative correlation of gene density with the LAD coverage across different chromosomes. This is in agreement with previously published reports that the chromosomal regions towards the nuclear periphery are gene-poor vs the gene-rich chromosomes toward the interior of the cell nucleus⁷⁰.

Our findings identify a subset of 5 and 6mer motifs in different representations (k-mer, forward reverse complement pair and motifs using variable positions) which had a higher preference of occurring in the LAD and the iLAD regions. These patterns of motif occurrences were observed to be conserved across all the chromosomes. We quantified these occurrences using our metric of LIR, where a ratio >1 suggests a higher preference for occurrences in the LAD regions. Alternatively, $LIR < 1$ signifies a higher preference for the occurrence of the motifs in the iLAD regions. The pattern of the conserved motifs suggested a base preference based on their occurrences in the LAD or the iLAD regions. A correlation of the base composition with the LIRs showed that The motifs that are present in the LADs have a higher preference toward AT composition. We observed a very strong negative correlation ($PCC \sim 0.9$) for both 5 and 6mer motifs. Our findings corroborate with other reports of AT preference in the LAD boundaries.

LADs can be facultative LADs or constitutive LADs based on the prevalence and conservation across different stages of the cell cycle. Our study focussed mostly on the constitutive LADS where the boundaries are conserved across different cell types and stages of differentiation during the cell cycle. Repeating our analysis on the genome wised lamina contacts across different cell types in mouse and *Drosophila* also suggested a conservation of the motif patterns evolutionarily. These findings highlight the importance of the conservation of the motif patterns for being recognized by their cognate binding partners thereby maintaining the transcriptional machinery of the cell.

Having established the boundaries of the LAD and the iLAD regions and identification of conserved motifs (both in LADs and iLADs) we further modelled a LAD-iLAD predictor based on the genome sequence. Using the motif occurrences information, we computed the EP sand TP from the training set of 7 randomly selected chromosomes and our model has optimal prediction with accuracy and MCC of 74% and 0.44 respectively. It is important to highlight the differences in the resolution of our prediction and that of experimental datasets. The HMM predicts a resolution of a motif window whereas the datasets available to use are of a much lower resolution of a few 100kbps. The model performs best in our shortlisted motifs (based on their preference of occurrences in the LAD and the iLAD regions). We obtained the shortlisted proteins

by finding interactors of Lamin B1 the and conservation of hydrogen bond donor-acceptor pattern of the query protein using JEDII. Based on consensus binding DNA 13 out of the 32 proteins showed complete overlap of our shortlisted DNA motifs that were used to design the HMM. Therefore, we are confident that this information will be crucial when identifying the protein-binding partners of these LAD regions. Existing tools to study genomes are mostly used to deny transitional activity peaks, specific binding domains of proteins or the euchromatin and heterochromatin states. To our knowledge, this is the first kind of predictor that is used to differentiate the genome into LAD and iLAD regions., which can be useful when studying chromosome organization. The novelty of the model is reading different regions of the genome at a resolution of how proteins would identify it.

Acknowledgements

Shortlisting protein targets using STRING-based searches and JEDII was done by Kshitij Bapat.

Chapter 7

Conclusions and future prospects

- [Summary of this study](#)
- [Compute database of gene network across multiple organisms](#)
- [Comparison of gene correlations across organisms](#)
- [Motif vector correlations to phylogeny](#)

7.1 Summary of this study:

In this study, we have read the genome at a resolution of motifs to understand the regulatory role that is encoded in the sequence information ([figure 7.1](#)). We established the optimal recognition of a DNA sequence by the cognate protein binding partners is 5-6 bases. This is established by studying the binding patch of 285 NR DNA protein complexes. The 5 and 6-mer motif distributions are obtained as the observed by expected (OE) ratio, which constitutes our motif vectors which were used to study the genome sequence at different resolutions. The motif distribution differed distinctly for the genome sequences across all the chromosomes against a randomly generated genome. This OE ratio when compared across the whole genome or a smaller stretch of the centromeric regions highlights a clustering within the chromosomes. Some chromosomes are more similar compared to others based only on the motif distribution. The motif vectors when obtained only in the centromeric regions can classify the chromosomes based on the type of the centromeres. The distribution differentiated the acrocentric chromosome (12,14,15, 21 and 22) from others.

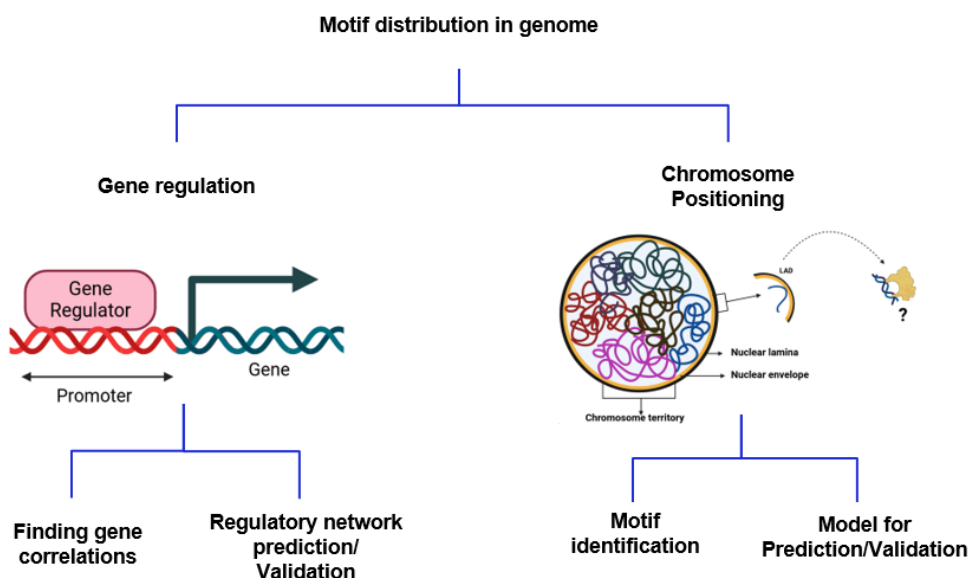


Figure 7.1: A schematic of the summary of the project

We compared the motif vectors for smaller regions of the genome like the gene promoter regions. We established a correlation score between all the annotated genes in the human genome, by comparing the OE ratios of all the motifs. These distributions of motifs were obtained for sizes 5 and 6-mers in regions 1kbps, 2kbps and 6kbps upstream of the gene transcription start site. The strong correlations between gene promoters give us insight into the shared regulatory mechanisms. The higher abundance of the motif consensus validates their recognition by the same transcription factor hence co-regulation. The high number of strong gene correlations (≥ 0.9) between chromosomes 13,14,15 21 and 22 also points towards their role in Robertsonian translocations. These higher structural rearrangements between the gene promoter patterns sharing high correlations can contribute to translocations.

We validated our correlations using three case studies. Firstly, we validated that genes across different chromosomes sharing high correlations also share common transcription factors. The high correlations are an attribute of the high abundance of the consensus binding motif of the common transcription factors. For the second validation study, we obtained the correlation score for the genes that are regulated by the AP1 transcription factor. We obtained that out of 19 genes that are reported to be regulated by the AP1 family of transcription factors 10 genes share a high correlation (≥ 0.5) among them. This suggesting correlation implies co-regulation in some cases. For the third case study, we established that high correlation also suggests spatial proximity. We validated our claims using Hi-C and pHi-C data between chromosomes 18 and 19. We observed that the regions of contacts between 18 and 19 that have genes in them have a enrichment of the highly correlated genes. Thus spatial conservation of gene promoters also suggests potential co-regulation.

We have built MAGNETIC (Motif Associated Gene NETwork in chromosomes) a webserver that gives the user access to all the correlations across all annotated genes. The webserver can be used to find gene associations or generate a network of multiple genes. These associations or networks are generated based on the distribution of motifs across the promoter regions of the genes. The correlations can be obtained for resolutions of 1kbps,2kbps and 6kbps upstream of the gene start site while reading a motif of size 5 or 6.

We have extended this analysis of reading geneic regions for understating gene regulation to chromosome positing. The nuclear boundary is made of the external

lining of the nuclear envelope. Beneath the nuclear envelope, there is a meshwork filamentous layer constituted by the nuclear lamina. The positioning of the chromosomes inside the nucleus is non-random that is linked to its association with the nuclear boundary. Certain regions in the chromosomes have a peripheral localization, whereas certain regions are located toward the interior. The gene density of the chromosomes and its expression depends on the localization of the chromosomal segments. Regions of the chromosome that make frequent contact with the nuclear lamina are referred to as the lamina associated domains (LADs). The regions interspersing them are termed as the inter-LADs (iLADs). We identified the motif distributions across the lamina associated domain positioning the chromosome. We found a conserved set of motifs that differentiates the LADs from the iLAD regions that are evolutionarily conserved across different organisms. On using this motif information, we designed a Hidden Markov Model-based predictor of LAD iLAD regions. The model differentiates the LAD from the iLAD regions reading the shortlisted motifs with an accuracy of ~75% and MCC of ~0.4. These shortlisted motifs also help us to identify the putative protein binding partners of the LADs.

7.2 Future prospects:

7.2.1 Compute database of gene network across multiple organisms:

The thesis was based on looking at the motif distributions across the different genes in the human genome. The study of motif distributions can be extended across different organisms. We have generated a database of these motif OE ratios and correlations between genes for 4 other organisms apart from the human genome. The future work will be to study the gene network across different organisms. Currently, we have computed the data for *C. elegans*, *Mus musculus*, *D. melanogaster* and *S. cerevisiae*. The future aim will be to design a pipeline to generate the motif distribution data and generate the correlations based on the motif vector across different genes across the genome. This pipeline will be linked to our MAGNETIC server so that the gene networks can be visualized within genomes and across genomes ([figure 7.2](#)).

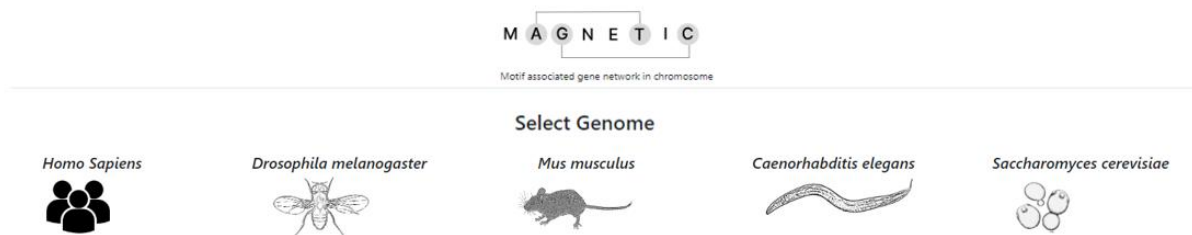


Figure 7.2: Future scope of MAGNETIC webserver incorporating multiple genomes

7.2.2 Comparison of gene correlations across organisms:

The thesis was based on looking at the motif distributions across the different genes in the human genome. As established previously the correlations between the genes were restricted to one genome. The future prospective of the project will be to extend this correlation to cross-genome homologues genes. Preliminary investigations of correlations of 13 housekeeping genes were done across humans and mouse. We selected housekeeping genes because these genes are involved in the essential functions of cell maintenance independent of the cell or the tissue type¹¹⁷. We obtained the gene correlations for 13 housekeeping genes in mouse¹¹⁸ and their respective human homologues ([table 7.1](#)). The correlations between these genes across organisms based on the motif OE in the 1kbps promoter region, range from -0.04 (for UBC genes) to 0.43 (for B2M genes).

Table 7.1: Correlation of motif vectors in 1kbps promoters of human vs mouse genes

Gene Name	Correlations
PGK1	-0.03
UBC	-0.04
TTR	0.29
TBP	-0.01
SDHA	-0.18
RPL27	0.29
RPL13A	0.25
OTX2	0.35
HPRT1	0.25

GAPDH	0.37
B2M	0.43
ATP5B	0.08
ACTB	0.34

Although the correlations in none of the 13 genes are ≥ 0.5 , there is an overlap of motifs that have the highest abundance in the gene promoters across both the organisms ([figure 7.3](#)). These motifs can be recognized by the same homologues protein binding partners across the organisms hence the higher repeats. Correlations across different organisms will help us in identifying such cases which need to be investigated further.

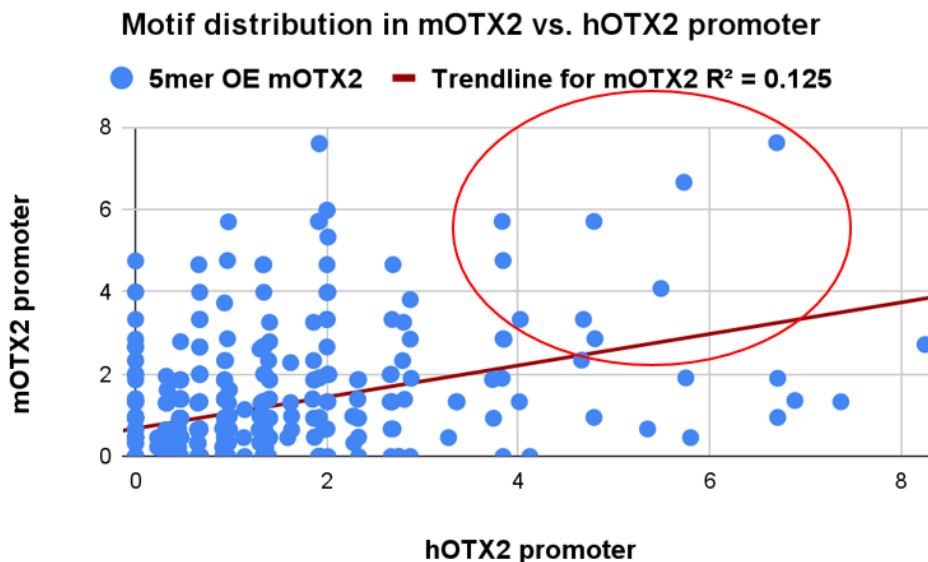


Figure 7.3: Comparison of motif OE in human vs mouse gene. The x and y axis shows the OE values for human and mouse OTX2 gene promoters respectively). The region in the scatterplot encircled in red highlights the common motifs having high OE ratios

7.2.3 Motif vector correlations to phylogeny:

We have established in the study that the abundance of motifs suggests a specific pattern of conservation at the level of motifs. This holds true when we are looking at smaller regions in the genome like the geneic regions or gene promoters as well as at

the level of the whole genome. Thus this method of correlation of motif vectors can be extended to study the evolutionary distance across genomes. A preliminary investigation using the motif correlations across 21 different strains of *S. cerevisiae*. The correlations between the strains (ranging from -1 to 1) were converted to distance by using $-1\log_{10}$. This distance score was used to generate the phylogenetic tree. We observed a trend of clustering of strains based on motif distributions across different chromosomes ([figure 7.4](#)). These results lead to further investigations on duplications of genome segments, how certain genome segments are translocation and the functional significance of the regions.

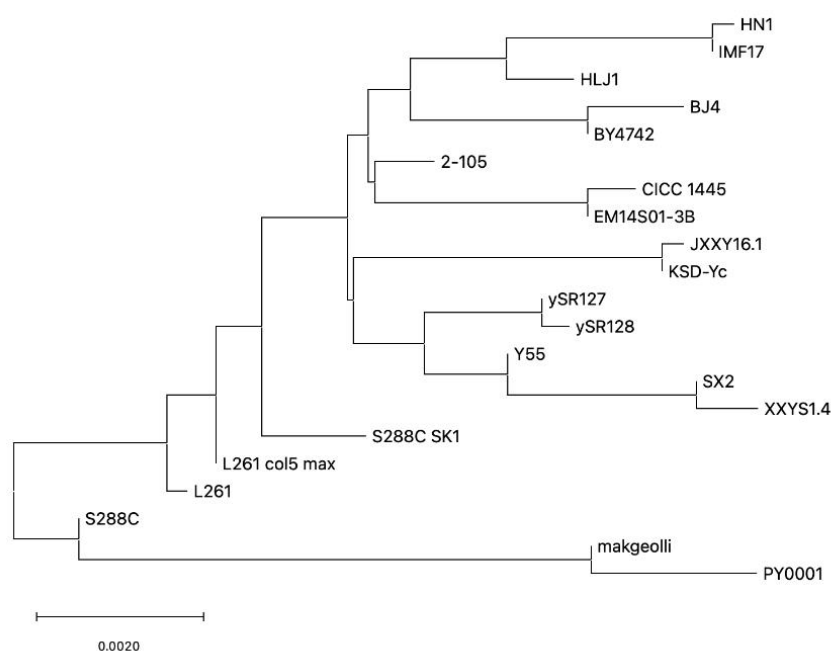


Figure 7.4: Phylogeny of different strains of yeast based on the motif distributions

Acknowledgements:

The analysis for motif vector correlations of the different strains of *S. cerevisiae* and the evolutionary tree was done by Aparajita Srinivasan.

Data availability

The motif OE values, correlation scores, and other supplementary information have been deposited in Zenodo at <https://doi.org/10.5281/zenodo.14055970>.

Publications

- Atreyi Chakraborty, Sumant Chopde, M.S Madhusudhan
bioRxiv 2024.09.18.613605; doi: <https://doi.org/10.1101/2024.09.18.613605>
- Atreyi Chakraborty*, Sumant Chopde, M.S Madhusudhan “Motif distribution in genomes gives insights into gene clustering and co-regulation”, *NAR* 2024; doi: <https://doi.org/10.1093/nar/gkae1178>
- Atreyi Chakraborty*, Ajinkya Ranshur, Rohan Topno, Saksham Shrimali, and M.S Madhusudhan MAGNETIC: A web server to find genome wide motif associated gene interaction network in chromosomes. (Manuscript submitted)
- Atreyi Chakraborty* and M.S Madhusudhan Insights in motif distribution in the chromosomal positioning in genome and prediction of lamina associated domains. (Manuscript in preparation)
- Sanjana Nair, Atreyi Chakraborty and M.S Madhusudhan. JEDII: Juxtaposition Enabled DNA-binding Interface Identifier (Manuscript in preparation)
- Sanjana Nair, Atreyi Chakraborty, Mrunal Pazare, M.S. Madhusudhan Hydrogen Bond Conservation Based Analysis of DNA-protein Specificity (Manuscript in preparation)

Bibliography

1. Alberts, B. *et al.* Studying Gene Expression and Function. (2002).
2. Nair, S. & Madhusudhan, M. S. JEDII: Juxtaposition Enabled DNA-binding Interface Identifier. *bioRxiv* 2022.05.19.492702 (2022) doi:10.1101/2022.05.19.492702.
3. Van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).
4. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).
5. Schoenfelder, S., Javierre, B. M., Furlan-Magaril, M., Wingett, S. W. & Fraser, P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. *J. Vis. Exp.* **2018**, 57320 (2018).
6. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10**, 402 (2009).
7. David Forney, G. Theory of electrodynamics in media in noninertial frames and applications. *Form. Struct. Elecfromagnetics. Amsterdam* **61**, 2589–2610 (1973).
8. Nguyen, M. N., Tan, K. P. & Madhusudhan, M. S. CLICK-topology-independent comparison of biomolecular 3D structures. doi:10.1093/nar/gkr393.
9. Nguyen, M. N. & Madhusudhan, M. S. Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.* **39**, (2011).
10. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science (80-.).* **376**, (2022).
11. Nurk, S. *et al.* The complete sequence of a human genome. *Science (80-.).* **376**, 44–53 (2022).
12. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
13. Bejerano, G., Haussler, D. & Blanchette, M. Into the heart of darkness: large-scale clustering of human non-coding DNA. *Bioinformatics* **20 Suppl 1**, (2004).
14. Guturu, H., Doxey, A. C., Wenger, A. M. & Bejerano, G. Structure-aided prediction of mammalian transcription factor complexes in conserved non-coding elements. *Philos.*

- Trans. R. Soc. Lond. B. Biol. Sci.* **368**, (2013).
15. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. *Science (80-.)*. **338**, 758–767 (2012).
 16. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.* **2021 121 12**, 1–14 (2021).
 17. Sémon, M. & Duret, L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. *Mol. Biol. Evol.* **23**, 1715–1723 (2006).
 18. Schneider, V. A. *et al.* EvSchneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, . *Genome Res.* **27**, 849–864 (2017).
 19. Church, D. M. A next-generation human genome sequence. *Science (80-.)*. **376**, 34–35 (2022).
 20. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science (80-.)*. **376**, (2022).
 21. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science (80-.)*. **376**, (2022).
 22. Cozzolino, F., Iacobucci, I., Monaco, V. & Monti, M. Protein-DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era. *J. Proteome Res.* **20**, 3018–3030 (2021).
 23. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164**, 1110–1121 (2016).
 24. Comfort, N. C. From controlling elements to transposons: Barbara McClintock and the Nobel Prize. *Trends Genet.* **17**, 475–478 (2001).
 25. Misteli, T. Higher-order genome organization in human disease. *Cold Spring Harb. Perspect. Biol.* **2**, (2010).
 26. Tjong, H. *et al.* Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1663–E1672 (2016).
 27. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping

- of nuclear lamina interactions. *Nature* **453**, 948–951 (2008).
28. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* **53**, 937–947 (1988).
 29. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem. Sci.* **25**, 99–104 (2000).
 30. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
 31. Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *Elife* **9**, 1–19 (2020).
 32. Nguyen, H. Q. *et al.* 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat. Methods* **2020 178** **17**, 822–832 (2020).
 33. Nir, G. *et al.* Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLOS Genet.* **14**, e1007872 (2018).
 34. Boninsegna, L. *et al.* Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations. *Nat. Methods* **2022 198** **19**, 938–949 (2022).
 35. Nix, D. A. & Eisen, M. B. GATA: A graphic alignment tool for comparative sequence analysis. *BMC Bioinformatics* **6**, 1–8 (2005).
 36. Fondon, J. W. & Garner, H. R. Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 18058–18063 (2004).
 37. Vinga, S. & Almeida, J. Alignment-free sequence comparison—a review. *Bioinformatics* **19**, 513–523 (2003).
 38. Reinert, G., Chew, D., Sun, F. & Waterman, M. S. Alignment-free sequence comparison (I): statistics and power. *J. Comput. Biol.* **16**, 1615–1634 (2009).
 39. Sarkar, B. K. *et al.* Determination of k-mer density in a DNA sequence and subsequent cluster formation algorithm based on the application of electronic filter. *Sci. Reports* **2021 111** **11**, 1–12 (2021).

40. Saw, A. K. *et al.* Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Sci. Rep.* **9**, (2019).
41. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nat.* **2009** 4617268 **461**, 1248–1253 (2009).
42. Chiu, T. P., Xin, B., Markarian, N., Wang, Y. & Rohs, R. TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.* **48**, D246–D255 (2020).
43. Yella, V. R. *et al.* Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res.* **46**, 11883–11897 (2018).
44. Ghoshdastidar, D. & Bansal, M. Flexibility of flanking DNA is a key determinant of transcription factor affinity for the core motif. *Biophys. J.* **121**, 3987–4000 (2022).
45. Homo sapiens genome assembly T2T-CHM13v2.0 - NCBI - NLM.
https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/.
46. NCBI Genome Data Viewer. <https://www.ncbi.nlm.nih.gov/genome/gdv/>.
47. Gonzalez, D. *et al.* ZNF143 protein is an important regulator of the myeloid transcription factor C/EBP α . *J. Biol. Chem.* **292**, 18924 (2017).
48. Humphray, S. J. *et al.* DNA sequence and analysis of human chromosome 9. *Nature* **429**, 369–374 (2004).
49. Sinclair, D. A. R., Schulze, S., Silva, E., Fitzpatrick, K. A. & Honda, B. M. Essential genes in autosomal heterochromatin of *Drosophila melanogaster*. *Genetica* **109**, 9–18 (2000).
50. Eberl, D. F., Duyf, B. J. & Hilliker, A. J. The role of heterochromatin in the expression of a heterochromatic gene, the rolled locus of *Drosophila melanogaster*. *Genetics* **134**, 277–292 (1993).
51. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344 (2023).
52. Stadhouders, R. *et al.* Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat. Genet.* **50**, 238 (2018).
53. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).

54. Casamassimi, A. & Ciccodicola, A. Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation. *Int. J. Mol. Sci.* **20**, (2019).
55. Hernandez-Garcia, C. M. & Finer, J. J. Identification and validation of promoters and cis-acting regulatory elements. *Plant Sci.* **217–218**, 109–119 (2014).
56. Reményi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* **11**, 812–815 (2004).
57. Groves, A. K. *et al.* Differential regulation of transcription factor gene expression and phenotypic markers in developing sympathetic neurons. *Development* **121**, 887–901 (1995).
58. Zhou, G. *et al.* NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* **47**, W234–W241 (2019).
59. Xia, J., Gill, E. E. & Hancock, R. E. W. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.* **10**, 823–844 (2015).
60. Spracklin, G. *et al.* Diverse silent chromatin states modulate genome compartmentalization and loop extrusion barriers. *Nat. Struct. Mol. Biol.* **2022 301 30**, 38–51 (2022).
61. Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **2015 61 6**, 1–7 (2015).
62. Hurst, L. D., Sachenkova, O., Daub, C., Forrest, A. R. R. & Huminiecki, L. A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol.* **15**, 1–26 (2014).
63. Saeed, S., Hassan, J., Javed, S. M., Shan, S. & Naz, M. A Familial Case of Robertsonian Translocation 13;14: Case Report. *Cureus* **14**, (2022).
64. Spinner, N. B., Conlin, L. K., Mulchandani, S. & Emanuel, B. S. Deletions and Other Structural Abnormalities of the Autosomes. *Emery Rimoin's Princ. Pract. Med. Genet.* 1–37 (2013) doi:10.1016/B978-0-12-383834-6.00051-3.
65. Harris, M. *et al.* Trisomy of Leukemic Cell Chromosomes 4 and 10 Identifies Children With B-Progenitor Cell Acute Lymphoblastic Leukemia With a Very Low Risk of Treatment Failure: A Pediatric Oncology Group Study. *Blood* **79**, 3316–3324 (1992).

66. Wong, K. F. & So, C. C. Acute myeloid leukemia with concomitant trisomies 4 and 10: a distinctive form of myeloid leukemia? *Cancer Genet. Cytogenet.* **127**, 74–76 (2001).
67. Goodship, J. *et al.* A submicroscopic translocation, t(4;10), responsible for recurrent Wolf-Hirschhorn syndrome identified by allele loss and fluorescent in situ hybridisation. *J. Med. Genet.* **29**, 451 (1992).
68. Vitalini, M. W., Dialynas, G., Wallrath, L. L., Mackey, S. R. & Stainbrook, S. C. Nucleus | Nuclear Organization, Chromatin Structure, and Gene Silencing. *Encycl. Biol. Chem. Third Ed.* **5**, 393–397 (2021).
69. Croft, J. A. *et al.* Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. *J. Cell Biol.* **145**, 1119 (1999).
70. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–220 (2001).
71. Mathelier, A. *et al.* JASPAR 2014: An extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, (2014).
72. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).
73. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-.).* **306**, 636–640 (2004).
74. de Souza, N. The ENCODE project. *Nat. Methods* **9**, 1046–1046 (2012).
75. A., M. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147.
76. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nat Cell Biol* **4**, E131–E136 (2002).
77. Zenz, R. & Wagner, E. F. Jun signalling in the epidermis: from developmental defects to psoriasis and skin tumors. *Int J Biochem Cell Biol* **38**, 1043–1049 (2006).
78. Zenz, R. *et al.* Activator protein 1 (Fos/Jun) functions in inflammatory bone and skin disease. *Arthritis Res. Ther.* **10**, 1–10 (2008).
79. Sefer, E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinforma. 2022 231* **23**, 1–39 (2022).

80. Nanni, L., Ceri, S. & Logie, C. Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. *Genome Biol.* **21**, 1–25 (2020).
81. Liu, Y. & Dekker, J. CTCF–CTCF loops and intra-TAD interactions show differential dependence on cohesin ring integrity. *Nat. Cell Biol.* **24**, 1516–1527 (2022).
82. Kuznetsova, M. A., Zaytseva, G. V., Zryachkin, N. I., Makarova, O. A. & Khmylevskaya, S. A. Patau Syndrome. *Vopr. Prakt. Pediatr.* **10**, 90–93 (2023).
83. Chung, W. K. Genetics. *Fetal and Neonatal Secrets* 266–289 (2014) doi:10.1016/B978-0-323-09139-8.00011-0.
84. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
85. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* (80-.). **376**, (2022).
86. Django documentation | Django documentation | Django. <https://docs.djangoproject.com/en/5.1/>.
87. Bruford, E. A. *et al.* HUGO Gene Nomenclature Committee (HGNC) recommendations for the designation of gene fusions. *Leuk.* **35**, 3040–3043 (2021).
88. Anshabo, A. T., Milne, R., Wang, S. & Albrecht, H. CDK9: A Comprehensive Review of Its Biology, and Its Role as a Potential Target for Anti-Cancer Agents. *Front. Oncol.* **11**, 678559 (2021).
89. Zhang, Q. *et al.* Roles and action mechanisms of WNT4 in cell differentiation and human diseases: a review. *Cell Death Discov.* **7**, 1–10 (2021).
90. Davidson, G. & Niehrs, C. Emerging links between CDK cell cycle regulators and Wnt signaling. *Trends Cell Biol.* **20**, 453–460 (2010).
91. Vitobello, A. *et al.* ADGRL1 haploinsufficiency causes a variable spectrum of neurodevelopmental disorders in humans and alters synaptic activity and behavior in a mouse model. *Am. J. Hum. Genet.* **109**, 1436–1457 (2022).
92. Malik, A. & Kanneganti, T. D. Function and regulation of IL-1 α in inflammatory diseases and cancer. *Immunol. Rev.* **281**, 124 (2018).
93. Xia, J., Benner, M. J. & Hancock, R. E. W. NetworkAnalyst - Integrative approaches

- for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res.* **42**, (2014).
94. Briand, N. & Collas, P. Lamina-associated domains: peripheral matters and internal affairs. *Genome Biol.* **21**, (2020).
 95. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harbor perspectives in biology* vol. 2 (2010).
 96. Gruenbaum, Y. *et al.* The nuclear lamina and its functions in the nucleus. *Int. Rev. Cytol.* **226**, 1–62 (2003).
 97. Dittmer, T. & Misteli, T. The lamin protein family. *Genome Biol.* **12**, 222 (2011).
 98. Vahabikashi, A., Adam, S. A., Medalia, O. & Goldman, R. D. Nuclear lamins: Structure and function in mechanobiology. *APL Bioeng.* **6**, 11503 (2022).
 99. Moir, R. D., Spann, T. P. & Goldman, R. D. The Dynamic Properties and Possible Functions of Nuclear Lamins. *Int. Rev. Cytol.* **162**, 141–182 (1996).
 100. Hozak, P., Sasseville, A. M. J., Raymond, Y. & Cook, P. R. Lamin proteins form an internal nucleoskeleton as well as a peripheral lamina in human cells. *J. Cell Sci.* **108**, 635–644 (1995).
 101. Dechat, T., Adam, S. A., Taimen, P., Shimi, T. & Goldman, R. D. Nuclear lamins. *Cold Spring Harb. Perspect. Biol.* **2**, (2010).
 102. Meaburn, K. J. *et al.* Primary laminopathy fibroblasts display altered genome organization and apoptosis. *Aging Cell* **6**, 139–153 (2007).
 103. Greil, F., Moorman, C. & van Steensel, B. [16] DamID: Mapping of In Vivo Protein–Genome Interactions Using Tethered DNA Adenine Methyltransferase. *Methods Enzymol.* **410**, 342–359 (2006).
 104. Aughey, G. N. & Southall, T. D. Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 25–37 (2016).
 105. Kind, J. *et al.* Single-cell dynamics of genome-nuclear lamina interactions. *Cell* **153**, 178–192 (2013).
 106. Kind, J. *et al.* Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134 (2015).
 107. Kind, J. *et al.* Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* **163**, 134–147 (2015).

108. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
109. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nat.* **596**, 583–589 (2021).
110. AlphaFold Protein Structure Database. <https://alphafold.ebi.ac.uk/>.
111. Nguyen, M. N., Tan, K. P. & Madhusudhan, M. S. CLICK--topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res.* **39**, (2011).
112. Sagendorf, J. M., Markarian, N., Berman, H. M. & Rohs, R. DNAProDB: An expanded database and web-based tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.* **48**, D277–D287 (2020).
113. Sagendorf, J. M., Berman, H. M. & Rohs, R. DNAProDB: An interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.* **45**, W89–W97 (2017).
114. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome – nuclear lamina interactions during differentiation. *Mol. Cell* **38**, 603 (2010).
115. van Bommel, J. G. *et al.* The Insulator Protein SU(HW) Fine-Tunes Nuclear Lamina Interactions of the Drosophila Genome. *PLoS One* **5**, 15013 (2010).
116. Lenain, C. *et al.* Massive reshaping of genome-nuclear lamina interactions during oncogene-induced senescence. *Genome Res.* **27**, 1634–1644 (2017).
117. Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J. & Lewis, N. E. What are housekeeping genes? *PLoS Comput. Biol.* **18**, (2022).
118. Ho, K. H. & Patrizi, A. Assessment of common housekeeping genes as reference for gene expression studies using RT-qPCR in mouse choroid plexus. *Sci. Reports* **11**, 1–15 (2021).

Copyright Form

Sign in/Register

Motif distribution in genomes gives insights into gene clustering and co-regulation

Author: Chakraborty, Atreyi; Chopde, Sumant
Publication: Nucleic Acids Research
Publisher: Oxford University Press
Date: 2024-12-09

Copyright © 2024, © The Author(s) 2024. Published by Oxford University Press on behalf of Nucleic Acids Research.

Welcome to RightsLink

Oxford University Press (OUP) has partnered with Copyright Clearance Center's RightsLink service to offer a variety of options for reusing this content.

Note: This article is available under the [Creative Commons CC-BY-NC](#) license and permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

For commercial reuse, permission must be requested below.

I would like to...

make a selection

*OUP offers a professional reprint service for high quality reproduction of articles from our scientific and medical journals. This service includes:

-English language reprints and ePrints
-Peer reviewed research or reviews
-Supplements/Collections of articles
-A professional high quality finish
-Journal style covers
-Prompt turnaround times and delivery directly to your office, warehouse or congress
Please contact our Corporate Sales team for a price quote: CorporateServices@oup.com

*Additionally, OUP offers a suite of licensing services to meet the needs of the healthcare industry worldwide.

For more information on how to obtain translated reprints/ ePrints, or Local Edition rights, please contact the Global Business Development & Rights team: translation.rights@oup.com

© 2025 Copyright - All Rights Reserved | [Copyright Clearance Center, Inc.](#) | [Privacy statement](#) | [Data Security and Privacy](#) | [For California Residents](#) | [Terms and Conditions](#)

Comments? We would like to hear from you. E-mail us at customer-care@copyright.com

133